BIOINFORMATICS TOOLS (AND WEB SERVER) FOR CANCER BIOMARKER DEVELOPMENT

EDITED BY: Xiangqian Guo, Liuyang Wang, Wan Zhu, Longxiang Xie and Jing Zhao

PUBLISHED IN: Frontiers in Oncology, Frontiers in Genetics and Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-261-6
DOI 10.3389/978-2-88966-261-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

BIOINFORMATICS TOOLS (AND WEB SERVER) FOR CANCER BIOMARKER DEVELOPMENT

Topic Editors:
Xiangqian Guo, Henan University, China
Liuyang Wang, Duke University, United States
Wan Zhu, Stanford University, United States
Longxiang Xie, Henan University, China
Jing Zhao, Chongqing Medical University, China

Citation: Guo, X., Wang, L., Zhu, W., Xie, L., Zhao, J., eds. (2020). Bioinformatics Tools (and Web Server) for Cancer Biomarker Development. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-261-6

Table of Contents

05 ***Editorial: Bioinformatics Tools (and Web Server) for Cancer Biomarker Development***
Longxiang Xie, Liuyang Wang, Wan Zhu, Jing Zhao and Xiangqian Guo

08 ***Identification of a Specific Gene Module for Predicting Prognosis in Glioblastoma Patients***
Xiangjun Tang, Pengfei Xu, Bin Wang, Jie Luo, Rui Fu, Kuanming Huang, Longjun Dai, Junti Lu, Gang Cao, Hao Peng, Li Zhang, Zhaohui Zhang and Qianxue Chen

19 ***Genome-Wide Profiling of Prognostic Alternative Splicing Pattern in Pancreatic Cancer***
Min Yu, Weifeng Hong, Shiye Ruan, Renguo Guan, Lei Tu, Bowen Huang, Baohua Hou, Zhixiang Jian, Liheng Ma and Haosheng Jin

32 ***Comprehensive Analysis of Expression and Prognostic Value of Sirtuins in Ovarian Cancer***
Xiaodan Sun, Shouhan Wang and Qingchang Li

46 ***Prognostic Roles of Central Carbon Metabolism–Associated Genes in Patients With Low-Grade Glioma***
Li Wang, Meng Guo, Kai Wang and Lei Zhang

58 ***A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis***
Guodong Yang, Yujiao Zhang and Jiyuan Yang

72 ***Comprehensive Analysis of Competitive Endogenous RNAs Network, Being Associated With Esophageal Squamous Cell Carcinoma and its Emerging Role in Head and Neck Squamous Cell Carcinoma***
Donghu Yu, Xiaolan Ruan, Jingyu Huang, Weidong Hu, Chen Chen, Yu Xu, Jinxuan Hou and Sheng Li

87 ***Prioritizing Gene Cascading Paths to Model Colorectal Cancer Through Engineered Organoids***
Yanyan Ping, Chaohan Xu, Liwen Xu, Gaoming Liao, Yao Zhou, Chunyu Deng, Yujia Lan, Fulong Yu, Jian Shi, Li Wang, Yun Xiao and Xia Li

101 ***Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis***
Hong Zheng, Guosen Zhang, Lu Zhang, Qiang Wang, Huimin Li, Yali Han, Longxiang Xie, Zhongyi Yan, Yongqiang Li, Yang An, Huan Dong, Wan Zhu and Xiangqian Guo

112 ***Identification of Core Gene Expression Signature and Key Pathways in Colorectal Cancer***
Xiang Ding, Houyu Duan and Hesheng Luo

125 ***OSgbm: An Online Consensus Survival Analysis Web Server for Glioblastoma***
Huan Dong, Qiang Wang, Ning Li, Jiajia Lv, Linna Ge, Mengsi Yang, Guosen Zhang, Yang An, Fengling Wang, Longxiang Xie, Yongqiang Li, Wan Zhu, Haiyu Zhang, Minghang Zhang and Xiangqian Guo

133 ***Integrated Analysis to Evaluate the Prognostic Value of Signature mRNAs in Glioblastoma Multiforme***

Ji'an Yang, Long Wang, Zhou Xu, Liquan Wu, Baohui Liu, Junmin Wang, Daofeng Tian, Xiaoxing Xiong and Qianxue Chen

142 ***Analysis of the Interaction Network of Hub miRNAs-Hub Genes, Being Involved in Idiopathic Pulmonary Fibers and its Emerging Role in Non-small Cell Lung Cancer***

Dong Hu Yu, Xiao-Lan Ruan, Jing-Yu Huang, Xiao-Ping Liu, Hao-Li Ma, Chen Chen, Wei-Dong Hu and Sheng Li

156 ***CD38 Predicts Favorable Prognosis by Enhancing Immune Infiltration and Antitumor Immunity in the Epithelial Ovarian Cancer Microenvironment***

Ying Zhu, Zhigang Zhang, Zhou Jiang, Yang Liu and Jianwei Zhou

169 ***OSluca: An Interactive Web Server to Evaluate Prognostic Biomarkers for Lung Cancer***

Zhongyi Yan, Qiang Wang, Zhendong Lu, Xiaoxiao Sun, Pengfei Song, Yifang Dang, Longxiang Xie, Lu Zhang, Yongqiang Li, Wan Zhu, Tiantian Xie, Jing Ma, Yijie Zhang and Xiangqian Guo

179 ***Single-Nucleotide Polymorphism Array Technique Generating Valuable Risk-Stratification Information for Patients With Myelodysplastic Syndromes***

Xia Xiao, Xiaoyuan He, Qing Li, Wei Zhang, Haibo Zhu, Weihong Yang, Yuming Li, Li Geng, Hui Liu, Lijuan Li, Huaquan Wang, Rong Fu, Mingfeng Zhao, Zhong Chen and Zonghong Shao

188 ***VisTCR: An Interactive Software for T Cell Repertoire Sequencing Data Analysis***

Qingshan Ni, Jianyang Zhang, Zihan Zheng, Gang Chen, Laura Christian, Juha Grönholm, Haili Yu, Daxue Zhou, Yuan Zhuang, Qi-Jing Li and Ying Wan

frontiers in Oncology

EDITORIAL
published: 20 October 2020
doi: 10.3389/fonc.2020.599085

Editorial: Bioinformatics Tools (and Web Server) for Cancer Biomarker Development

Longxiang Xie[1], *Liuyang Wang*[2], *Wan Zhu*[3], *Jing Zhao*[4] *and Xiangqian Guo*[1]*

[1] Cell Signal Transduction Laboratory, Department of Preventive Medicine, Bioinformatics Center, Henan Provincial Engineering Center for Tumor Molecular Medicine, Institute of Biomedical Informatics, School of Basic Medical Sciences, Henan University, Kaifeng, China, [2] Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, Durham, NC, United States, [3] Department of Anesthesia, Stanford University, Stanford, CA, United States, [4] Department of Pathophysiology, Chongqing Medical University, Chongqing, China

Keywords: bioinformatics, webserver, prognostic, biomarker, TCGA, GEO, RNA sequence

Editorial on the Research Topic

Bioinformatics Tools (and Web Server) for Cancer Biomarker Development

Cancer remains a severe public health burden globally. The identification of molecular biomarkers play significant roles in diagnosis, treatment and prognosis of human cancers (1). Up to now, the tumor molecular heterogeneity and lack of sufficient biomarkers are two of the major difficulties in cancer treatment and prognostication. With the advance of recent development of high-throughput microarray and sequencing technologies, the public cancer transcriptomic databases, including The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), have increased dramatically (2). These databases offer additional resources and opportunities for biomarker discovery and validation (2). Unfortunately, those resources are not efficiently explored, and translation of stored high dimension data into clinical use are not feasible for clinicians and basic researchers without much bioinformatics background. Therefore, the user-friendly online web servers/tools are urgently needed for researchers. In this Research Topic, we have collected a series of original research articles and reviews, providing a number of useful web resources and tools. Those tools will facilitate better and accurate discovery of cancer biomarkers and expedite their clinical translation.

Currently, several powerful bioinformatics webservers/tools, such as KM plotter, GEPIA (Gene Expression Profiling Interactive Analysis), Oncomine and TIMER (Tumor Immune Estimation Resource), have been developed to analyze the public transcriptomic datasets along with clinical information for oncology research (3–6). However, limitations are still present for these webservers/tools, such as tedious registration process or single data source. To overcome these limitations, Yan et al. developed a new survival analysis web-server OSluca for lung cancer based on 5,245 clinical samples from TCGA, GEO and Roepman study. With OSluca, the users are able to assess the prognostic value of gene of interest, and the results will be presented by Kaplan-Meier (KM) plot, Hazard ratio (HR), and log-rank *p*-value. Dong et al. also collected 684 samples with long-term follow-up clinical information from 7 TCGA, GEO and Chinese Glioma Genome Atlas (CGGA) datasets, and developed a survival analysis online tool OSgbm for glioblastoma. In recent years, T cell repertoire sequencing (TCRSeq) data have been rapidly developed, however, tools for comprehensive analysis and visualization of TCR-Seq data have not been developed. Ni et al. developed a tool called VisTCR (Visual TCRSeq), an interactive software with a graphical user interface (GUI) for TCR data management, short-read sequence mapping, and post-analysis of TCR clonotype. VisTCR can be used to perform clonotype extraction and downstream analyses within a single data management framework, which will greatly help TCRseq data management and

OPEN ACCESS

Edited and reviewed by:
Claudio Sette,
Catholic University of the Sacred Heart, Italy

***Correspondence:**
Xiangqian Guo
xqguo@henu.edu.cn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *26 August 2020*
Accepted: *11 September 2020*
Published: *20 October 2020*

Citation:
Xie L, Wang L, Zhu W, Zhao J and Guo X (2020) Editorial: Bioinformatics Tools (and Web Server) for Cancer Biomarker Development. Front. Oncol. 10:599085. doi: 10.3389/fonc.2020.599085

analysis in cancer immunotherapy. In a review of webserver/tools for cancer prognosis analysis, Zheng et al. described 22 webservers/tools for survival analysis based on mRNA, ncRNA, DNA and protein data, including LOGpc, KM plotter, GEPIA, OncoLnc, TCPA, MethSurv, PrognoScan, SurvExpress, and UALCAN, and they also gave a detailed description of the software usage, characteristics and algorithms of all these tools. They also discussed several major challenges and future directions in this area.

Those online webservers/tools for survival analysis would help clinician and researchers to discover novel prognostic biomarkers (3–6), to find the important therapeutic targets, and to investigate the potential molecular mechanisms of tumorigenesis and progression. Using a series of online databases, such as Oncomine and GEPIA, Kaplan-Meier plotter, TCGA, and cBioPortal, Sun et al. systematically analyzed the expression variation and prognostic value of sirtuins (SIRTs) 1–7 in ovarian cancer. The bioinformatics analysis showed that SIRT1-4, 6 and 7 may be novel prognostic biomarkers. Zhu et al. used a range of online tools, including Oncomine, GEPIA, TISIDB, and Kaplan-Meier plotter, to evaluate the expression and prognostic value of *CD38*. The results showed that compared with normal ovarian tissue, *CD38* is highly expressed in epithelial ovarian cancer (EOC), and higher *CD38* expression is associated with better prognosis. In addition, *CD38* was found to be associated with tumor-infiltrating lymphocytes (TILs), especially with activated CD8C T cells by TIMER. This implies the vital immunoregulatory role of *CD38* in the EOC microenvironment, and provides a novel prognostic biomarker and potential immunotherapy target. Yu et al. assembled 45,313 pancreatic cancer-specific AS (Alternative splicing) events of 10,623 genes from the TCGA and SpliceSeq database, and performed the cox univariate analyses of overall survival (OS). They found 6,711 AS events are remarkably associated with OS in pancreatic cancer. Notably, AS events of five genes including *DAZAP1*, *RBM4*, *ESRP1*, *QKI*, and *SF1*, were found to be significantly correlated with OS. Using the DriverDBv2, 13 driver genes were identified correlated with survival-associated AS events, including *TP53* and *CDC27*. These findings uncover that the aberrant AS patterns might serve as prognostic predictors in pancreatic cancer. Ding et al. performed the comprehensive characterization of differentially expressed genes between 65 normal colon tissues and 74 CRC samples, and identified 20 hub genes with a high degree of connectivity from the protein–protein interaction (PPI) network. Furthermore, knockdown of one hub gene, *MAD2L1*, significantly inhibited the CRC cell growth by impairing cell cycle progression and inducing cell apoptosis, implying that *MAD2L1* could be as a novel potential biomarker for diagnosis and therapy in CRC.

Single nucleotide polymorphism array (SNP-A) detects population-level genomic polymorphisms and chromosomal abnormalities such as submicroscopic or cryptic deletions or duplications (7). Xiao et al. used SNP-A technique to investigate the chromosomal abnormalities in 350 myelodysplastic syndromes (MDSs) patients and 26 healthy individuals. They showed that chromosomal aberrations contributed to a unfavorable prognosis in patients with myelodysplastic syndromes, and were closely related with an increased risk of transformation to typical myelodysplastic syndrome in patients with idiopathic cytopenia of undetermined significance. Thus, SNP-A can help assess the prognosis of patients with MDSs and the risk of disease progression for patients with ICUS.

Engineered organoids with sequential introducing driver mutations can provide important new clues for studying the mechanisms of cancer progression. Ping et al. developed an comprehensive strategy to capture the dynamic progression of CRC and prioritize gene cascading paths to model CRC through engineered organoids. From the single-mutant to quintuple-mutant engineered organoids, they characterized the functional activities of hallmark signatures and filled the substantial biological gaps between the engineered organoids and the CRC samples.

Although many single-gene cancer biomarkers have been reported, multi-gene signatures capture more information and may be more powerful for cancer prognosis, and they can be developed by analyzing public microarray data and RNA sequencing data (8). Based on the TCGA database and weighted gene co-expression network analysis (WGCNA), Tang et al. used Kaplan-Meier survival analysis and multivariate Cox regression method, and identified a four-gene prognostic signature (*CLEC5A*, *FMOD*, *FKBP9*, *LGALS8*) that was related with OS and recurrence time of 524 GBM patients. Those signature genes divided GBM patients into high-risk and low-risk groups, and the 5-years survival rate of the low-risk group was significantly higher than that of the high-risk group. Yang et al. profiled 4 GEO datasets and TCGA dataset from GBM patients, and performed the differential expression analysis, WGCNA and Cox regression analysis to identify core genes associated with clinical outcomes. A four-gene prognostic signature (*SLC12A5*, *CCL2*, *IGFBP2*, and *PDPN*) that was able to divide GBM patients into high-risk and low-risk groups. High-risk group showed higher mortality than low risk group by Kaplan–Meier curve. Yang et al. obtained 502 differential expressed miRNAs based on miRNA expression profiles of CRC patients from TCGA. Among these miRNAs, a novel five-miRNA signature (hsa-miR-5091, hsamiR-10b-3p, hsa-miR-9-5p, hsa-miR-187-3p, hsa-miR-32-5p) that could predict OS of CRC patients was constructed, verified and assessed in training group, testing group, and entire cohort. Furthermore, univariate and multivariate cox regression analysis showed that the five-miRNA signature could serve as an independent prognostic factor in CRC. Wang et al. investigated the expression profile of 63 central carbon metabolism–associated genes in 514 diffuse low-grade glioma cases (astrocytoma, oligodendroglioma, and oligoastrocytoma) from TCGA, and explored the prognostic roles of individual genes and the multiple-gene combination by Kaplan–Meier curve and multivariate cox regression analysis. The results showed that a four genes-signature (*RAF1*, *AKT3*, *IDH1*, and *FGFR1*) is positively associated with OS in patients with astrocytoma, suggesting that multigene expression signature is able to predict the prognosis of low-grade glioma patients.

Increasing studies have demonstrated that the competitive endogenous RNAs (ceRNA) regulation network plays an important role in cancer development (9). Yu et al. used WGCNA to construct the lncRNA co-expression networks,

miRNA co-expression networks, and mRNA co-expression networks based on TCGA-ESCC RNAseq data. They identified 21 hub lncRNAs, seven hub miRNAs, and nine hub mRNAs, and constructed a ceRNA network, the similar ceRNA network was also built for head and neck squamous cell carcinoma (HNSCC) by using UALCAN, OncomiR and OncoLnc webtools. Two hub genes including *TBC1D2* and *ATP6V0E1* were found to be associated with the survival time of HNSCC. The ceRNAs network might provide common mechanisms involving in ESCC and HNSCC. The same group also constructed the gene co-expression networks and miRNA co-expression networks in Idiopathic pulmonary fibrosis (IPF) based on two GEO datasets (GSE3257 and GSE3258), then validated the clinical significance of the genes and the miRNAs in other three GEO datasets (GSE10667, GSE70866, and GSE27430). They identified seven hub miRNAs and six hub mRNAs, and constructed an interaction network of hub miRNAs-hub genes, which was also analyzed in non-small cell lung cancer (NSCLC). In addition, six hub genes and three miRNAs were found to be associated with the survival time of lung adenocarcinoma (LUAD).

The increasing multi-omics data greatly help us to understand cancer biology and identification of molecular biomarkers, but add additional layers of difficulty in data processing and analyses. In this special issue, a range of powerful bioinformatics tools/webservers for data analysis have been developed, and they will easily assist clinical and basic science researchers in biomarker development and validation. Of note, the bioinformatics tools/web servers presented here still need lots of improvements, for example, integrating the tumor tissue image, multi-omics network mapping, multi-gene signature assessment, and nomogram construction. After tackling these problems in future, the bioinformatics tools/webservers will be more powerfully for discovering cancer biomarkers and innovative cancer therapies.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This study was supported by National Natural Science Foundation of China (No. 81602362), Program for Innovative Talents of Science and Technology in Henan Province (No. 18HASTIT048), and supporting grant of Bioinformatics Center of Henan University (Nos. 2018YLJC01 and 2019YLXKJC04).

REFERENCES

1. Zhou C, Zhong X, Song Y, Shi J, Wu Z, Guo Z et al. Prognostic biomarkers for gastric cancer: an umbrella review of the evidence. *Front oncol.* (2019) 9:1321. doi: 10.3389/fonc.2019.01321
2. Wang X, Hu S, Ji W, Tang Y, Zhang S. Identification of genes associated with clinicopathological features of colorectal cancer. *J Int Med Res.* (2020) 48:0300060520912139. doi: 10.1177/0300060520912139
3. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* (2017) 77:e108–e110. doi: 10.1158/0008-5472.CAN-17-0307
4. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia.* (2004) 6:1. doi: 10.1016/S1476-5586(04)80047-2
5. Gyorffy B, Surowiak P, Budczies J, Lanczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS ONE.* (2013) 8:e82241. doi: 10.1371/journal.pone.0082241
6. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* (2017) 45:98–102. doi: 10.1093/nar/gkx247
7. Arenillas L, Mallo M, Ramos F, Guinta K, Barragán E, Lumbreras E, et al. Single nucleotide polymorphism array karyotyping: a diagnostic and prognostic tool in myelodysplastic syndromes with unsuccessful conventional cytogenetic testing. *Genes Chromosomes Cancer.* (2013) 52:1167–1177. doi: 10.1002/gcc.22112
8. Cai L, Wang F, Zhang L, Wang Q, Guo X. Systematic review of prognostic gene signature in gastric cancer patients. *Front Bioeng Biotech.* (2020) 8:805. doi: 10.3389/fbioe.2020.00805
9. Sanchez-Mejias A, Tay Y. Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics. *J Hematol Oncol.* (2015) 8:30. doi: 10.1186/s13045-015-0129-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fonc.2019.00812

Identification of a Specific Gene Module for Predicting Prognosis in Glioblastoma Patients

Xiangjun Tang[1,2,3†], Pengfei Xu[1†], Bin Wang[2], Jie Luo[2], Rui Fu[2], Kuanming Huang[2], Longjun Dai[2], Junti Lu[2], Gang Cao[2], Hao Peng[2], Li Zhang[2], Zhaohui Zhang[4*] and Qianxue Chen[1*]*

[1] Department of Neurosurgery, Renmin Hospital of Wuhan University, Wuhan, China, [2] Department of Neurosurgery, Taihe Hospital, Hubei University of Medicine, Shiyan, China, [3] Department of Neurosurgery, Affiliated Hospital of Xi'an Jiaotong University Health Science Center, Xi'an, China, [4] Department of Neurology, Renmin Hospital of Wuhan University, Wuhan, China

Introduction: Glioblastoma (GBM) is the most common and malignant variant of intrinsic glial brain tumors. The poor prognosis of GBM has not significantly improved despite the development of innovative diagnostic methods and new therapies. Therefore, further understanding the molecular mechanism that underlies the aggressive behavior of GBM and the identification of appropriate prognostic markers and therapeutic targets is necessary to allow early diagnosis, to develop appropriate therapies and to improve prognoses.

Methods: We used a weighted gene co-expression network analysis (WGCNA) to construct a gene co-expression network with 524 glioblastoma samples from The Cancer Genome Atlas (TCGA). A risk score was then constructed based on four module genes and the patients' overall survival (OS) rate. The prognostic and predictive accuracy of the risk score were verified in the GSE16011 cohort and the REMBRANDT cohort.

Results: We identified a gene module (the green module) related to prognosis. Then, multivariate Cox analysis was performed on 4 hub genes to construct a Cox proportional hazards regression model from 524 glioblastoma patients. A risk score for predicting survival time was calculated with the following formula based on the top four genes in the green module: risk score = $(0.00889 \times EXP_{CLEC5A}) + (0.0681 \times EXP_{FMOD}) + (0.1724 \times EXP_{FKBP9}) + (0.1557 \times EXP_{LGALS8})$. The 5-year survival rate of the high-risk group (survival rate: 2.7%, 95% CI: 1.2–6.3%) was significantly lower than that of the low-risk group (survival rate: 8.8%, 95% CI: 5.5–14.1%).

Conclusions: This study demonstrated the potential application of a WGCNA-based gene prognostic model for predicting the survival outcome of glioblastoma patients.

Keywords: glioblastoma, WGCNA, prognostic model, cox proportional hazards regression model, nomogram

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Zhitong Bing,
Lanzhou University, China
Xinyu Chen,
Stanford University, United States

****Correspondence:***
Li Zhang
zhanglith@163.com
Zhaohui Zhang
zhzhqing1990@163.com
Qianxue Chen
chenqx666@whu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *19 April 2019*
Accepted: *08 August 2019*
Published: *27 August 2019*

Citation:
Tang X, Xu P, Wang B, Luo J, Fu R, Huang K, Dai L, Lu J, Cao G, Peng H, Zhang L, Zhang Z and Chen Q (2019) Identification of a Specific Gene Module for Predicting Prognosis in Glioblastoma Patients. Front. Oncol. 9:812. doi: 10.3389/fonc.2019.00812

INTRODUCTION

Glioma is one of the most common types of malignant brain tumors and has a very poor prognosis (1). The efficacy of conventional surgery plus radio- and chemotherapy is poor. Several signature molecular markers have been used in the diagnosis, therapy and prognosis of glioma. For example, methyl guanine methyl transferase (MGMT) promoter methylation is considered a

predictive marker for the resistance of glioblastoma (GBM) to chemotherapy with temozolomide (2). The 1p/19q co-deletion is a molecular signature of oligodendroglial tumors and a predictive marker for the response of anaplastic gliomas to vincristine (PCV) chemotherapy. High WT-1 expression is significantly associated with worse outcomes in diffuse astrocytic tumors. IDH1/IDH2 mutations have a strong favorable prognostic value across all glioma histopathological grades (3–5). With the advancement of gene technology, molecular signatures for the classification of gliomas have become prominent in recent years. The 2016 revision of the World Health Organization (WHO) classification of tumors of the central nervous system (6) includes novel classes of diffuse gliomas based on genomic features. Though molecular diagnostics increase diagnostic accuracy and prognostic yield compared to previous histology-based classifications, the current clinical prediction and treatment outcomes are still not satisfactory (7). As GBM is notoriously heterogeneous and complex, multi-parameter markers are much more accurate for cancer prognosis than a single biomarker. Therefore, a proper analytical model is highly desirable.

In the present study, we identified gene modules related to the overall survival (OS) and recurrence time of GBM based on The Cancer Genome Atlas (TCGA) database and weighted gene co-expression network analysis (WGCNA). The TCGA database contains genomic expression, sequence, methylation, and copy number variation data on over 11,000 individuals and over 30 kinds of cancers (8, 9). WGCNA is based on a system of biological methods for describing the correlation patterns among genes and modules of highly correlated genes. By using Kaplan-Meier survival analysis and multivariate Cox regression analysis, we identified a prognostic model for GBM patients based on gene characteristics. Our findings may provide novel insight toward developing a promising predictive tool for the prognosis of GBM.

MATERIALS AND METHODS

Patients

A total of 906 glioma cases were collected from three databases in this study, including 528 samples from TCGA (https://portal.gdc.cancer.gov), 219 samples from REMBRANDT (https://gdoc.georgetown.edu/gdoc/), and 159 samples from the GSE16011 dataset (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16011). Forty-six samples were excluded due to a lack of OS information. As shown in **Figure 1**, we grouped cases from TCGA into a training cohort, whereas all cases from REMBRANDT and GSE16011 were used for validation.

Abbreviations: WGCNA, Weighted Gene Co-expression Network Analysis; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; GS, gene significance; MS, module significance; EGFR, epidermal growth factor receptor; HR, hazard ratio; EXP, expression; MAPK, mitogen-activated protein kinase; CLEC5A, C-type lectin member 5A; FMOD, fibromodulin; FKBP, FK506-binding proteins; LGALS8, lectin, galactose binding, soluble 8; CIMP:CpG island methylator phenotype; IDH1, Isocitratedehydrogenase 1; IDH2, isocitrate dehydrogenase 2; MGMT, O6-methylguanine-DNA methyltransferase; FPKM, fragments per kilobase per million; TOM, topological overlap measure; GO, gene ontology; CC, cellular component; MF, molecular function; BP, biological process; KEGG, Kyoto Encyclopedia of Genes and Genomes; ROC, receiver operating characteristic curve; AUC, area under the curve; JEV, Japanese Encephalitis Virus; AIC, akaike information criterion.

Data Pre-processing

Microarray data of the 906 samples were normalized by the affy package. All data were filtered to reduce outliers. For genes with several probes, the median of all probes was chosen. For probes with missing values, the impute package (http://bioconductor.org/packages/release/bioc/html/impute) was used to fill the missing values. Finally, 12,700 genes were obtained from the TCGA dataset.

Construction of the Weighted Gene Co-expression Network

By choosing 6 as a soft threshold, a weighted gene co-expression network was constructed using the R package WGCNA (10), which has the approximate scale-free fundamental property of the biological gene networks. A co-expression similarity matrix was composed of the absolute value of the correlation between the expression levels of transcripts. The network modules were generated using the topological overlap measure (TOM) (11), and the dynamic hybrid cut method (a bottom-up algorithm) was used to identify co-expression gene modules (12). Finally, the modules with highly correlated genes were merged, and the minimum height for merging modules was set to 0.2. Gene significance (GS) and module significance (MS) were calculated to measure the correlation between the sample traits (recurrence time, CpG island methylator phenotype (CIMP) status, survival time, status, IDH1 status, MGMT status, subtype, age and sex) of either the genes or modules. The targeted module genes were visualized with Cytoscape 3.5.1 software (13).

Functional Enrichment Analysis

The biological process (BP) ontology of the modules was analyzed by Gene Ontology (GO) (14), while pathway enrichment was analyzed by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (15). The function of module genes was verified by the R package clusterProfiler (16). The corrected P-value (false discovery rate, FDR) < 0.05 was identified as a significant outcome.

Identification of the Predicted Survival of Glioblastoma Patients by the Cox Proportional Hazards Regression Model

To verify the significance of the genes screened above, the 436 green module genes were first screened using univariate Cox proportional hazards regression, and the 230 genes with p-value <0.05 was selected for the advanced analysis (**Supplemental Data 2**). According to the p-value, we selected only the top 14 survival-related genes for visualization using the R package forestplot. Then, a multivariate Cox regression model analysis was performed to establish a Cox proportional hazards regression prognostic model, which was calculated as follows: risk score $= \Sigma(\mathrm{C} \times \mathrm{EXPgene})$, where EXP was the mRNA expression of the crucial gene, and C was the regression coefficient for the corresponding gene in the multivariate Cox hazard model analysis. The optimal model was determined based on akaike

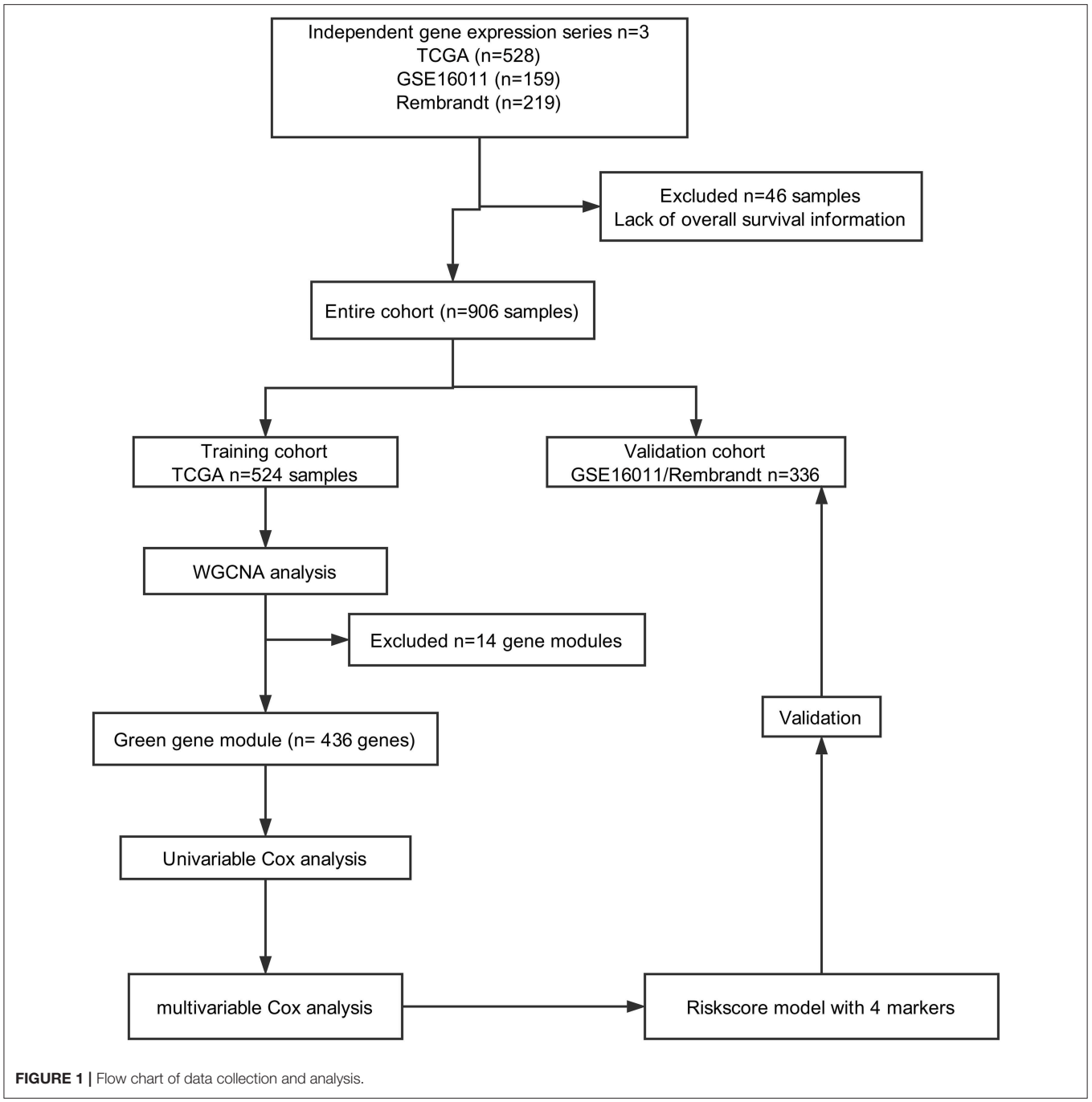


FIGURE 1 | Flow chart of data collection and analysis.

information criterion (AIC). The relevant codes were provided in the **Supplemental File**. The samples were divided into a high-risk group and a low-risk group according to the median risk score of the training dataset from TCGA.

Statistical Analysis

Survival curves were constructed by the Kaplan-Meier method and compared by the log-rank test, which was carried out through the R package survival. The sensitivity and specificity of the survival prediction based on the risk score were depicted by a time-dependent receiver operating characteristic (ROC) curve using the R package survivalROC. Gene set enrichment analysis (GSEA) was used to identify the pathways that were significantly enriched between the high- and low-risk groups. The Cox regression model was used to perform the multivariable survival analysis and generate nomograms. Calibration curves were used to assess whether the actual outcomes approximately predicted outcomes for the nomogram. Nomogram and calibration curves were performed with the rms package (https://CRAN.R-project.org/package=rms). The discrimination of the nomogram was

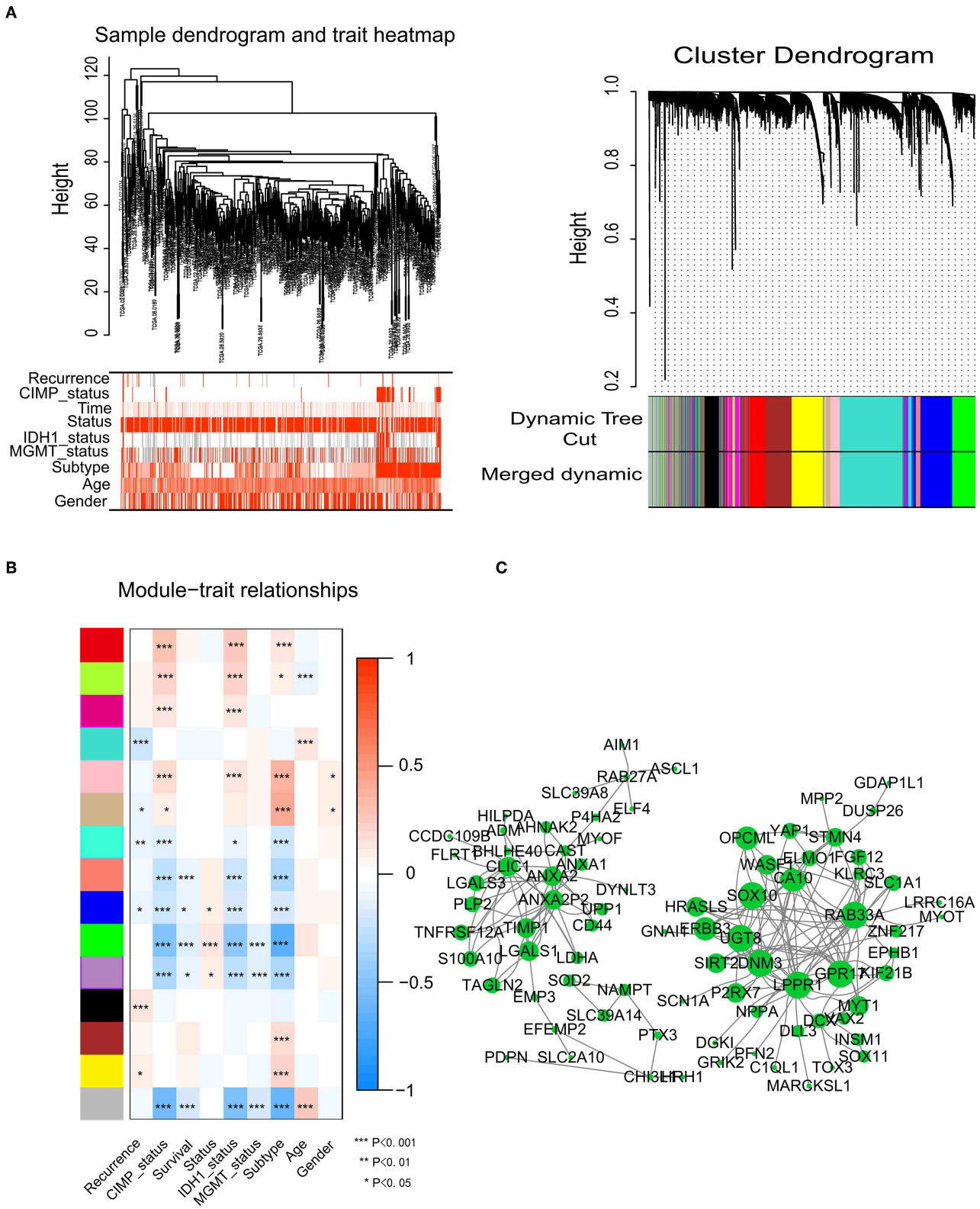


FIGURE 2 | Network construction of the weighted co-expressed genes and their associations with clinical traits. **(A)** Hierarchical clustering tree of the TCGA-GBM samples based on the training cohort. Dendrogram tips are labeled with the TCGA-GBM unique name. In the hierarchical dendrogram, lower branches correspond to

(Continued)

FIGURE 2 | higher co-expression. The branches of the cluster dendrogram correspond to the 15 different gene modules based on topological overlaps. Each piece of the leaves on the cluster dendrogram represents a gene. **(B)** Module-trait relationships. The background colors of the numbers represent the strength of the correlation between the gene module and the clinical traits, which increased from blue to red. Each column corresponds to a clinical trait. **(C)** Visualization of the co-expression network of the green module. The larger the nodes and the numerous edges, the more significant the gene is. Based on weight, not all genes were represented.

measured and compared by the C-index. All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant. Statistical analyses were conducted using R software (version 3.4.3, www.r-project.org).

RESULTS

Pre-processing of RNA Sequence Data and Clinical Data

In total, 906 glioblastoma microarray and clinical data were downloaded from TCGA, REMBRANDT and GSE16011. We constructed an mRNA expression matrix with gene symbols and patient barcodes. Furthermore, outlier samples with expression quantities <20% were screened. A total of 46 samples were discarded owing to the lack of OS information. Finally, the top 5,000 genes with the greatest variance obtained from the training cohort were used in the WGCNA studies.

Identification of Modules Associated With Glioma Survival Status

To identify significant gene modules, we constructed a gene co-expression network with WGCNA. With a scale-free network and topological overlaps, we generated a hierarchical clustering tree based on the dynamic hybrid cut (**Figure 2A**). Finally, 15 gene modules were identified, and the branches of the tree represent different gene modules. The non-co-expressed genes were included in the "gray" module, which was not further analyzed (**Figure 2B**). The relationships of the fifteen modules were analyzed with clinical traits, such as survival time, recurrence time, age, and sex. The green module correlated significantly with survival status (**Figure 2B**). A total of 436 genes were included in the green module.

Visualization of Green Module Genes

Network screening was used to detect the hub genes in the green module. The co-expression network of the green module was visualized with a Cytoscape graph. As shown in **Figure 2C**, the hub genes were centrally located in the modules and may be the key elements of the modules. The larger the nodes and the numbers of the edges, the more significant the gene is. When depicted based on weight, not all genes were represented.

Functional Enrichment Analysis

We performed a functional enrichment analysis of the green module using GO analysis. As shown in **Figures 3A–D**, enriched BPs were mainly involved in the positive regulation of cellular component biogenesis. The cellular components (CCs) were mainly enriched in focal adhesion and the cell substrate adherens junction. Enriched molecular functions (MFs) were mainly involved in cell adhesion molecule binding. KEGG pathway analysis showed that the MAPK signaling pathway was the most enriched pathway, followed by proteoglycans in cancer and the regulation of the actin cytoskeleton. The results suggested that these genes were closely related to cell adhesion function.

Identification and Validation of a Cox Proportional Hazards Regression Model

We further selected all genes of the green module to perform a univariate Cox analysis (**Figure 3E**). Then, multivariate Cox analysis was performed on the four genes that were significantly related to survival time. A Cox proportional hazards regression model was constructed with the TCGA cohort. The risk score for predicting survival time was calculated with the following formula based on the four genes: risk score $= (0.00889 \times \mathrm{EXP}_{\mathrm{CLEC5A}}) + (0.0681 \times \mathrm{EXP}_{\mathrm{FMOD}}) + (0.1724 \times \mathrm{EXP}_{\mathrm{FKBP9}}) + (0.1557 \times \mathrm{EXP}_{\mathrm{LGALS8}})$.

We divided patients from the training set into high-risk (n = 262) and low-risk (n = 262) groups according to the median of the risk score. The 1- and 3-year areas under the ROC curve were 0.62 and 0.71, respectively, indicating a high predictive value. Additionally, the predictive model can function as a good predictive indicator of the survival of glioma patients, which was confirmed by Kaplan-Meier curves. Patients with high-risk scores exhibited worse OS according to the Kaplan-Meier curves. The 5-year and 3-year survival rates of the high-risk group (2.7 and 6.8%, respectively) were significantly worse than those of the low-risk group (8.8 and 18.9%, respectively; **Figure 4A**). Moreover, the Kaplan-Meier curves confirmed that the four genes could function as predictive indicators for the survival of GBM patients in the training cohort (**Figures 3F–I**).

Furthermore, we assessed the prognostic effect of different clinical characteristics using a univariate Cox proportional hazards regression model. The results showed that CIMP status, IDH1 status, MGMT status, age, and risk score were associated with OS ($P < 0.01$) (**Table 1**). However, the multivariate regression model showed that the risk score and age were independent prognostic factors associated with OS.

To confirm that the proposed risk score model has similar prognostic value in different populations, the same formula was applied to the GSE16011 and REMBRANDT cohorts. The results showed that patients in the high-risk group had a significantly lower OS rate than those in the low-risk group in both the GSE16011 and REMBRANDT cohorts (**Figures 4B–C**). The functional GSEA showed that the high-risk group was highly enriched in genes closely related to base excision repair, the cell cycle, DNA replication, and ribosome function (**Figure 5A**).

Construction of a Predictive Nomogram

To develop a quantitative method to predict patients' OS rate, we constructed a nomogram in the TCGA cohort. The risk score was stratified into high- and low-risk groups based on the

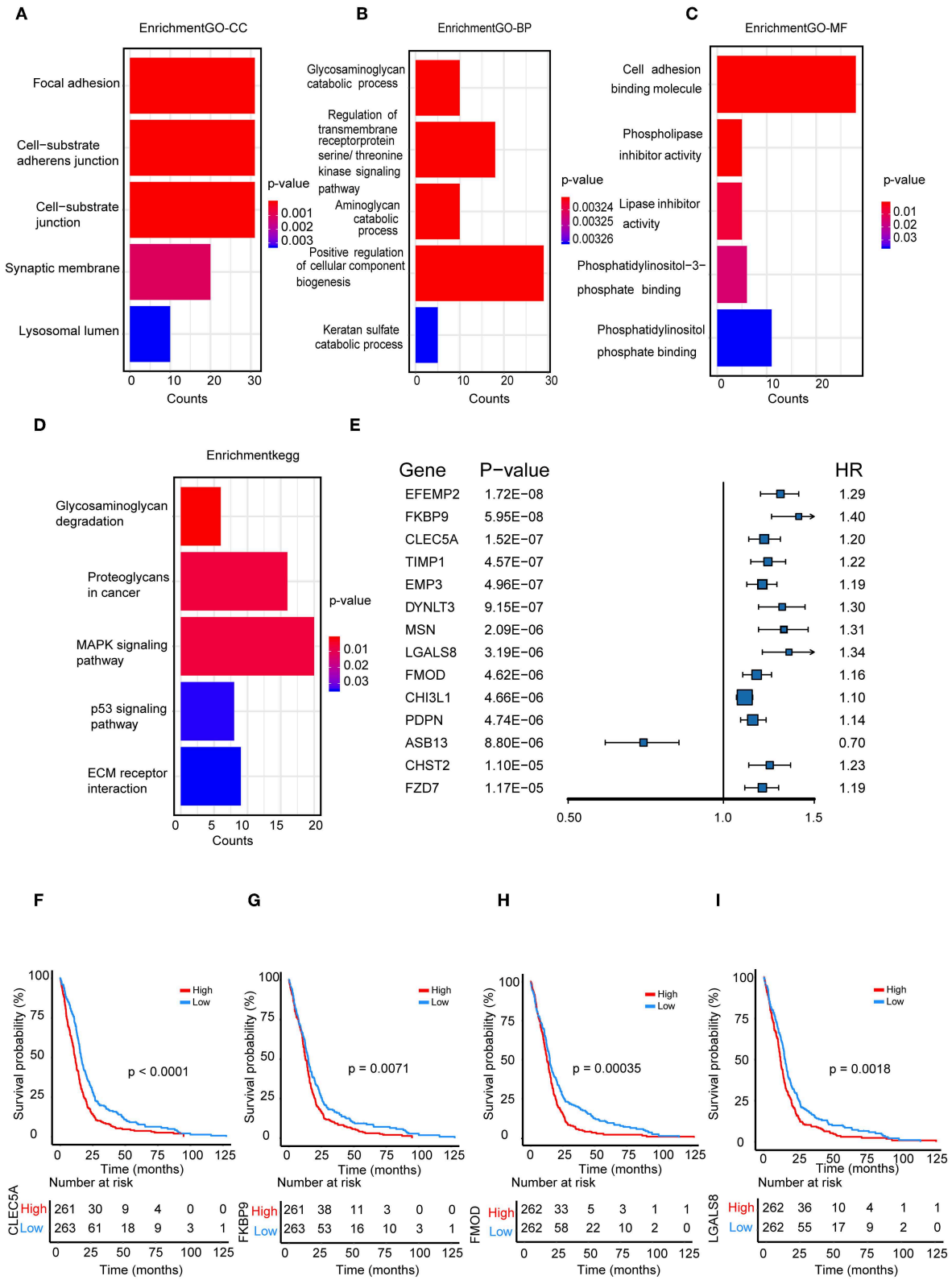


FIGURE 3 | Functional enrichment analysis. **(A)** Biological process **(B)** cellular component, **(C)** molecular function; **(D)** enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis for hub genes related to survival time; **(E)** The top 14 genes which were significantly related to survival time in univariate analysis; **(F–I)** Kaplan-Meier curves for CLEC5A,FKBP9, FMOD, and LGALS8 in the TCGA cohort.

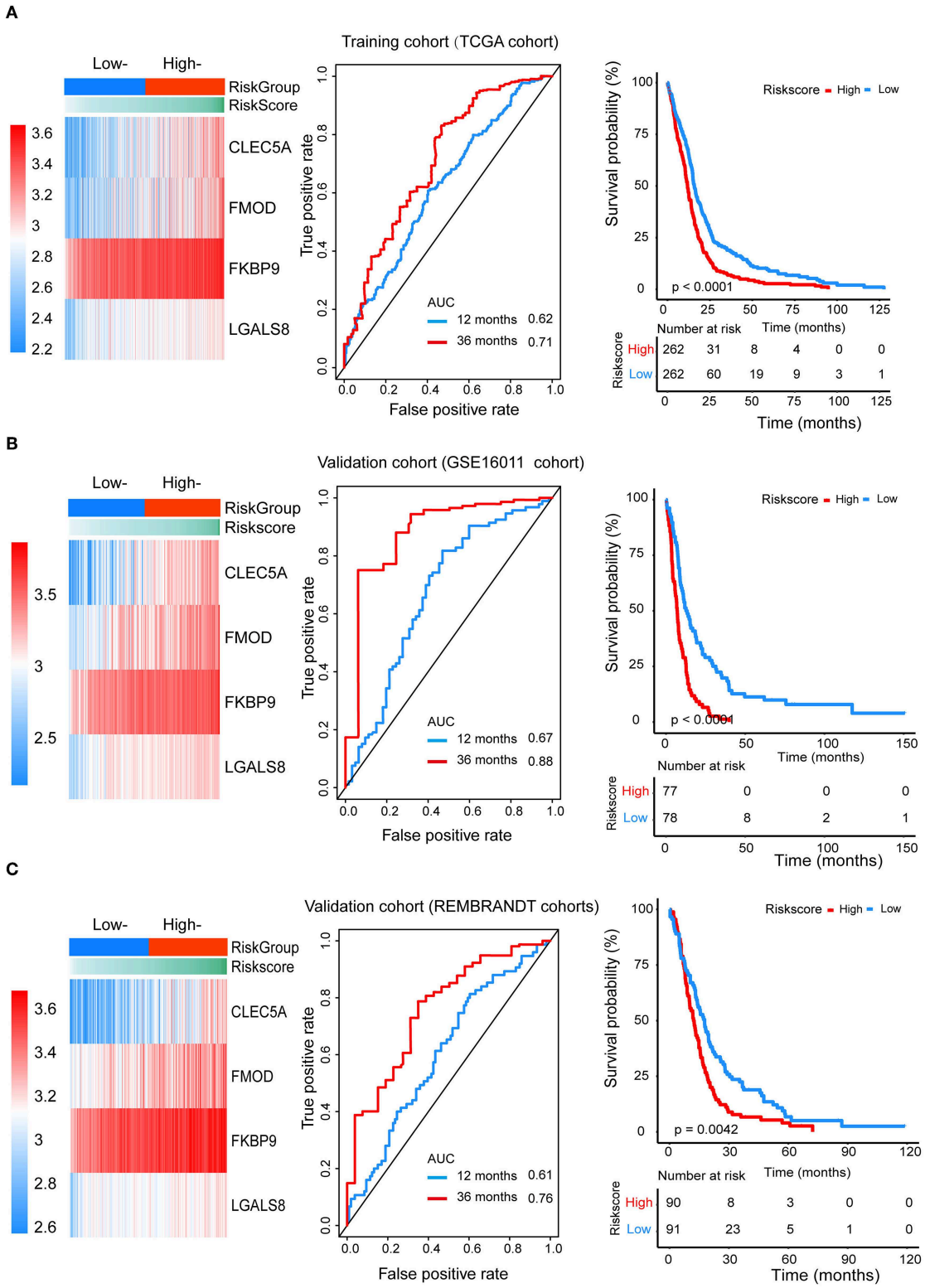


FIGURE 4 | The prognostic efficiency of the Cox proportional hazards regression model. Heat map of the model genes in **(A)** training set of the TCGA, **(B)** test set of GSE16001, **(C)** test set of Rembrandt; ROC curves of the four genes signature for predicting 12- and 36-months survival of glioblastoma. The 12- and 36-months areas (AUC) under the ROC curves indicate higher predictive value; Kaplan–Meier curves analyze the survival of the high-risk group and the low-risk group, the high-risk group had the worse outcome ($P < 0.001$).

TABLE 1 | The prognostic effect of different clinical characteristics.

	Univariate analysis[a]			Multivariate analysis[b]		
	HR	**95%CI**	***P*-value**	**HR**	**95%CI**	***P*-value**
CIMP-status	0.35	0.24–0.5	<0.001	0.29	0.04–2.19	0.232
IDH1-status	0.34	0.21–0.55	<0.001	1.8	0.23–14.08	0.573
MGMT-status	0.69	0.54–0.87	<0.001	0.84	0.64–1.1	0.205
Subtype	0.93	0.86–1.01	0.07	-	-	-
Age	1.03	1.03–1.04	<0.001	1.03	1.02–1.04	<0.001
Gender	1.16	0.96–1.41	0.13	-	-	-
Risk score	1.57	1.3–1.89	<0.001	1.49	1.14–1.94	0.003

[a] *These data were used to perform the Cox proportional hazards regression.*
[b] *Multivariate analysis used stepwise addition of clinical covariates related to survival in univariate analysis ($P < 0.01$) and the ultimate models contained those covariates that were significantly associated with survival ($P < 0.01$).*

median. The predictors included age, risk group, and IDH1 status (**Figure 5B**). Due to the lack of IDH1 mutation information in the REMBRANDT cohorts, the calibration curves for the 1- and 3-year OS rates were well-predicted in only the TCGA and GSE16011 cohorts (C-index: 0.65 for the TCGA cohort and 0.68 for the GSE16011 cohort; **Figures 5C,D**).

DISCUSSION

Gliomas are the most common and malignant brain tumors with poor prognosis, especially GBM. The most promising treatments, such as surgery, radiation, and chemotherapy with temozolomide, improve survival measured in only weeks rather than years (17). Precise studies of GBM biology and molecular markers have renewed our understanding of GBM. In 2008, Parsons et al. first proposed subtypes of GBM based on specific gene alterations (18). In 2016, the WHO revised the classification of tumors of the central nervous system based on gene technology and molecular signatures. The classification contained some well-known biomarkers, such as MGMT methylation, 1p/19q co-deletion, IDH 1 or 2, and EGFR. Recently, Suchorska et al. reported that amino acid positron emission tomography (PET)-based metabolic imaging can be used as a promising tool for the non-invasive characterization of molecular features and to provide additional prognostic information (19). These classifications and studies helped with prognosis, survival time, and response to treatment. As GBMs are heterogeneous and complex, molecular signatures are superior to single biomarkers in the prognosis of glioma.

To identify a gene signature associated with the survival status of GBM patients, we first constructed a weighted gene co-expression network in 524 glioma samples and generated the survival time-specific green module. The detected hub genes in the green module were significantly correlated with the survival status of patients with GBM. The GO and KEGG functional enrichment analysis showed that the genes that were closely related to adhesion function, adhesion molecules and the MAPK signaling pathway accounted for the highest proportion of green module genes. Adhesion function is a key factor in glioma invasiveness, and adhesion molecules play an important role in gliomagenesis. The MAPK pathway regulates the activity of transcription factors that function in proliferation, survival, differentiation, and apoptosis (20). Furthermore, this signaling pathway is also activated by EGFR signaling. The MAPK pathway could also be directly or indirectly activated through mutations of downstream components. In high-grade gliomas, MAPK-activated samples presented prolonged survival in comparison to other high-grade tumors. In low-grade gliomas, the presence of activated MAPK was also a predictor of favorable patient outcome, regardless of fusion or hotspot mutation events (21).

To analyze the relationship between survival time and the hub genes of the green module, we selected 436 genes for univariate Cox analysis. Our survival analysis by constructing a Cox proportional hazards regression model showed that CLEC5A, FMOD, FKBP9, and LGALS8 were highly associated with OS. CLEC5A/MDL-1 is a member of the myeloid C-type lectin family expressed in macrophages and neutrophils, which is strongly associated with the activation and differentiation of myeloid cells and has been implicated in the progression of multiple acute and chronic inflammatory diseases. Research by Batliner et al. suggested that CLEC5A/MDL-1 could activate a signaling cascade that results in the activation of downstream kinases in inflammatory responses (22) and maintain lesional macrophage survival, causing their accumulation (23). Another report showed that Japanese encephalitis virus (JEV) directly interacted with CLEC5A. Additionally, anti-CLEC5A mAb could repair the blood-brain barrier, attenuate neuroinflammation, and protect mice from JEV-induced lethality (24). Recently, R. Chai reported that CLEC5A was also a prognostic biomarker of GBM (25). FKBP9 is a peptidyl–prolyl isomerase and is a member of this protein family. It has been implicated in neurodegeneration, mainly through accelerating fibrillization (26, 27). Fibromodulin (FMOD), as a GBM-upregulated gene, promotes glioma cell migration through its ability to generate the formation of filamentous actin stress fibers. FMOD-induced glioma cell migration is dependent on the integrin-FAK-Src-Rho-ROCK signaling pathway (28). FMOD was also reported to be a prognostic biomarker in GBM (29). LGALS8 plays functional roles in promoting GBM cell proliferation and clonal sphere formation (30). Though CLEC5A and FKBP9 have not been reported in glioma-related studies, their features play important roles in cell metabolism and pathological processes. Further studies are needed to explore their relationship with glioma. Therefore, CLEC5A, FMOD, FKBP9, and LGALS8 could be considered crucial prognostic factors in the OS of glioma patients.

In this study, we constructed a prognostic score model of a four-gene signature. The univariate Cox proportional hazards regression result demonstrated that this four-gene signature, together with CIMP status, IDH1 status, MGMT status, and age, was highly associated with OS. The independent prognostic significance was also verified according to a multivariate regression model. The ability of the four-gene model to predict survival outcomes was further confirmed by the validation cohorts from the REMBRANDT and GSE16011 datasets. To

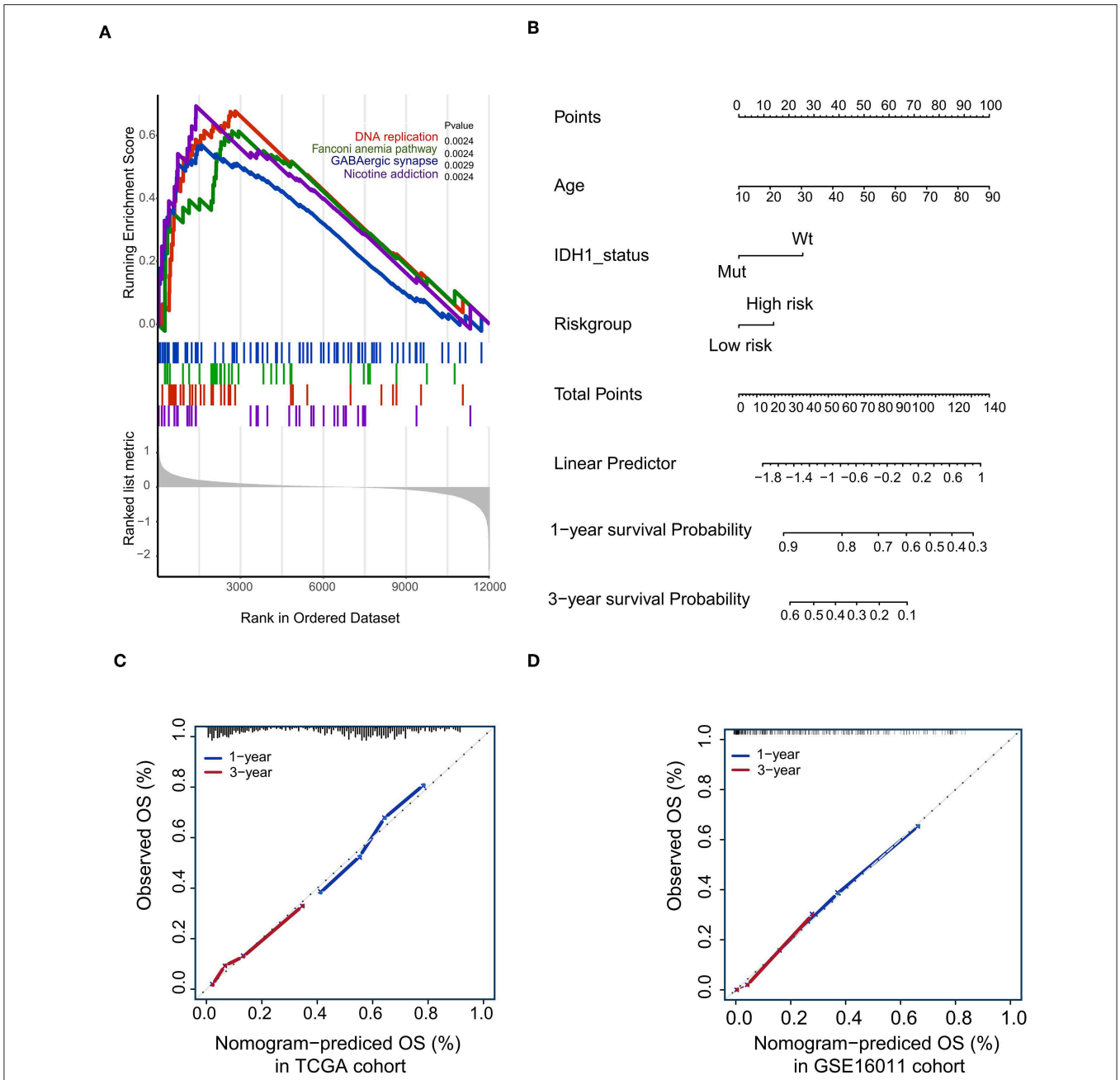


FIGURE 5 | Gene-set enrichment analysis (GSEA) and Nomogram. **(A)** The GSEA showed that high-risk group highly enriched in Base excision repair, Cell cycle, DNA replication, Ribosome; **(B)** Nomogram to predict the 1- and 3-year OS. Calibration curve for OS nomogram model in the TCGA cohort **(C)** and GSE16011 cohort **(D)**.

further strengthen the accuracy of the model, we combined age, IDH1 status, and risk group to fit a Cox proportional regression model in the TCGA cohort and used a nomogram for visualization. The calibration curves showed high predictive ability in the TCGA and GSE16011 cohorts. Our analysis showed that the four-gene model is likely a promising and viable prognostic signature for the survival status of glioma patients.

In summary, through the construction of a gene co-expression network with data from the TCGA database, a green module with a survival signature was identified using the WGCNA approach. The hub genes were selected from the green module genes and visualized with Cytoscape. By constructing a Cox proportional hazards regression model, four genes were finally identified and used in univariate and multivariate Cox analyses, thereby composing a four-gene module with the risk score = $(0.00889 \times EXP_{CLEC5A}) + (0.0681 \times EXP_{FMOD}) + (0.1724 \times EXP_{FKBP9}) + (0.1557 \times EXP_{LGALS8})$. This four-gene module represents a promising and viable prognostic signature for the

survival outcome of GBM patients. The present study revealed the potential application of a WGCNA-based gene prognostic model for predicting the survival outcomes of GBM patients.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov.

ETHICS STATEMENT

Ethical approval was waived since we used only publicly available data and materials in this study.

AUTHOR CONTRIBUTIONS

XT, PX, and LZ: conception and design. XT, PX, BW, and JLuo: acquisition of data. XT, PX, RF, KH, and ZZ: analysis and interpretation of data. XT, LD, and ZZ: writing and review of the manuscript. JLu, GC, HP, LZ, and QC: study supervision. All authors have read and approved the final version of this manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (No. 81702482), Natural Science Foundation of Hubei Province of China (No. 2017CFB562).

ACKNOWLEDGMENTS

We gratefully acknowledge The Cancer Genome Atlas pilot project (established by NCI and NHGRI), which made the genomic data and clinical data of glioma available.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.00812/full#supplementary-material

REFERENCES

1. Goodenberger ML, Jenkins RB. Genetics of adult glioma. *Cancer Genet-NY*. (2012) 205:613–21. doi: 10.1016/j.cancergen.2012.10.009
2. Kessler T, Sahm F, Sadik A, Stichel D, Hertenstein A, Reifenberger G, et al. Molecular differences in IDH wildtype glioblastoma according to MGMT promoter methylation. *Neuro Oncol*. (2018) 20:367–79. doi: 10.1093/neuonc/nox160
3. Westphal M, Lamszus K. Circulating biomarkers for gliomas. *Nat Rev Neurol*. (2015) 11:556–66. doi: 10.1038/nrneurol.2015.171
4. Schwab DE, Lepski G, Borchers C, Trautmann K, Paulsen F, Schittenhelm J. Immunohistochemical comparative analysis of GFAP, MAP - 2, NOGO - A, OLIG - 2 and WT - 1 expression in WHO 2016 classified neuroepithelial tumors and their prognostic value. *Pathol Res Pract*. (2018) 214:15–24. doi: 10.1016/j.prp.2017.12.009
5. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. (2010) 17:98–110. doi: 10.1016/j.ccr.2009.12.020
6. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol*. (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1
7. Zhang C, Cheng W, Ren X, Wang Z, Liu X, Li G, et al. Tumor purity as an underlying key factor in Glioma. *Clin Cancer Res*. (2017) 23:6279–91. doi: 10.1158/1078-0432.CCR-16-2598
8. Wang Z, Jensen MA, Zenklusen JC. A practical guide to The Cancer Genome Atlas (TCGA). *Methods Mol Biol*. (2016) 1418:111–41. doi: 10.1007/978-1-4939-3578-9_6
9. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. (2015) 19:A68–77. doi: 10.5114/wo.2014.47136
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform*. (2008) 9:559. doi: 10.1186/1471-2105-9-559
11. Li A, Horvath S. Network module detection: affinity search technique with the multi-node topological overlap measure. *BMC Res Notes*. (2009) 2:142. doi: 10.1186/1756-0500-2-142
12. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for Bioinformatics R. *Bioinformatics*. (2008) 24:719–20. doi: 10.1093/bioinformatics/btm563
13. Diaz-Montana JJ, Gomez-Vela F, Diaz-Diaz N. GNC-app: A new Cytoscape app to rate gene networks biological coherence using gene-gene indirect relationships. *Biosystems*. (2018) 166:61–5. doi: 10.1016/j.biosystems.2018.01.007
14. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. (2004) 32:D258–61. doi: 10.1093/nar/gkh036
15. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. (2000) 28:27–30. doi: 10.1093/nar/28.1.27
16. Yu G, Wang L, Han Y, He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. (2012) 16:284–7. doi: 10.1089/omi.2011.0118
17. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn M, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. (2005) 352:987–96. doi: 10.1056/NEJMoa043330
18. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma Multiforme. *Science*. (2008) 321:1807–12. doi: 10.1126/science.1164382
19. Suchorska B, Albert NL, Bauer EK, Tonn J, Galldiks N. The role of amino-acid PET in the light of the new WHO classification 2016 for brain tumors. *Q J Nucl Med Mol Im*. (2018) 62:267–71. doi: 10.23736/S1824-4785.18.03090-X
20. Aldape K, Zadeh G, Mansouri S, Reifenberger G, von Deimling A. Glioblastoma: pathology, molecular mechanisms and markers. *Acta Neuropathol*. (2015) 129:829–48. doi: 10.1007/s00401-015-1432-1
21. Ryall S, Krishnatry R, Arnoldo A, Buczkowicz P, Mistry M, Siddaway R, et al. Targeted detection of genetic alterations reveal the prognostic impact of H3K27M and MAPK pathway aberrations in paediatric thalamic glioma. *Acta Neuropathol Commun*. (2016) 4:93. doi: 10.1186/s40478-016-0353-0
22. Batliner J, Mancarelli MM, Jenal M, Reddy VA, Fey MF, Torbett BE, et al. CLEC5A (MDL-1) is a novel PU.1 transcriptional target during myeloid differentiation. *Mol Immunol*. (2011) 48:714–9. doi: 10.1016/j.molimm.2010.10.016
23. Xiong W, Wang H, Lu L, Xi R, Wang F, Gu G, et al. The macrophage C-type lectin receptor CLEC5A (MDL-1) expression is associated with early plaque progression and promotes macrophage survival. *J Transl Med*. (2017) 15:234. doi: 10.1186/s12967-017-1336-z

24. Chen S, Liu R, Wu M, Lin Y, Chen S, Tan DT, et al. CLEC5A Regulates Japanese Encephalitis Virus-Induced Neuroinflammation and Lethality. *PLoS Pathog.* (2012) 8:e10026554. doi: 10.1371/journal.ppat.1002655
25. Chai R, Zhang K, Wang K, Li G, Huang R, Zhao Z, et al. A novel gene signature based on five glioblastoma stem-like cell relevant genes predicts the survival of primary glioblastoma. *J Cancer Res Clin Oncol*. (2018) 144:439–47. doi: 10.1007/s00432-017-2572-6
26. Gerard M, Deleersnijder A, Demeulemeester J, Debyser Z, Baekelandt V. Unraveling the role of peptidyl-prolyl isomerases in neurodegeneration. *Mol Neurobiol*. (2011) 44:13–27. doi: 10.1007/s12035-011-8184-2
27. Deleersnijder A, Van Rompuy A, Desender L, Pottel H, Buee L, Debyser Z, et al. Comparative analysis of different Peptidyl-Prolyl isomerases reveals FK506-binding Protein 12 as the most potent enhancer of alpha-Synuclein Aggregation. *J Biol Chem*. (2011) 286:26687–701. doi: 10.1074/jbc.M110.182303
28. Mondal B, Patil V, Shwetha SD, Sravani K, Hegde AS, Arivazhagan A, et al. Integrative functional genomic analysis identifies epigenetically regulated fibromodulin as an essential gene for glioma cell migration. *Oncogene*. (2017) 36:71–83. doi: 10.1038/onc.2016.176
29. Xiong J, Bing Z, Su Y, Deng D, Peng X. An integrated mRNA and microRNA expression signature for glioblastoma multiforme prognosis. *PLoS ONE*. (2014) 9:e98419. doi: 10.1371/journal.pone.0098419
30. Laks DR, Crisman TJ, Shih MY, Mottahedeh J, Gao F, Sperry J, et al. Large-scale assessment of the gliomasphere model system. *Neuro Oncol*. (2016) 18:1367–78. doi: 10.1093/neuonc/now045

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fonc.2019.00773

Genome-Wide Profiling of Prognostic Alternative Splicing Pattern in Pancreatic Cancer

Min Yu[1†‡], Weifeng Hong[2†], Shiye Ruan[1,3†], Renguo Guan[1,3†‡], Lei Tu[1], Bowen Huang[1,3], Baohua Hou[1], Zhixiang Jian[1], Liheng Ma[2] and Haosheng Jin[1*]*

[1] Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China, [2] Department of Medical Imaging, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [3] The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Yan-feng Gao,
Zhengzhou University, China
Zhenyu Shi,
Henan University, China

***Correspondence:**
Min Yu
yumin@gdph.org.cn
Haosheng Jin
thundercry@163.com

[†]These authors have contributed equally to this work

[‡]**ORCID:**
Min Yu
orcid.org/0000-0003-1875-740X
Renguo Guan
orcid.org/0000-0002-9487-7369

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 24 April 2019
Accepted: 31 July 2019
Published: 27 August 2019

Citation:
Yu M, Hong W, Ruan S, Guan R, Tu L, Huang B, Hou B, Jian Z, Ma L and Jin H (2019) Genome-Wide Profiling of Prognostic Alternative Splicing Pattern in Pancreatic Cancer. Front. Oncol. 9:773. doi: 10.3389/fonc.2019.00773

Alternative splicing (AS) has a critical role in tumor progression and prognosis. Our study aimed to investigate pancreatic cancer-specific AS events using RNA-seq data, gaining systematic insights into potential prognostic predictors. We downloaded 10,623 genes with 45,313 pancreatic cancer-specific AS events from the Cancer Genome Atlas (TCGA) and SpliceSeq database. Cox univariate analyses of overall survival suggested there was a remarkable association between 6,711 AS events and overall survival in pancreatic cancer patients ($P < 0.05$). The area under the curves (AUC) of the receiver operator characteristic curves (ROC) of risk score was 0.89 for final prognostic predictor. Results indicated that AS events of DAZAP1, RBM4, ESRP1, QKI, and SF1 were significantly associated with overall survival. The results of FunRich showed that transcription factors KLF7, GABPA, and SP1 were the most highly related to survival-associated AS genes. Furthermore, using DriverDBv2, we identified 13 driver genes associated with survival-associated AS events, including TP53 and CDC27. Thus, we concluded that the aberrant AS patterns in pancreatic cancer patients might serve as prognostic predictors.

Keywords: alternative splicing, TCGA, pancreatic cancer, prognosis, driver gene

INTRODUCTION

During the pre-mRNA splicing, introns are removed, and the exons are left to form the final mRNA products. In this process, exons which are left vary, and thus, one single gene may generate multiple mRNA isoforms by alternative splicing (AS). More than 95% of human genes undergo AS, and most of them vary in levels across different cells and tissues (1). Variations in AS may result in a spectrum of consequences from completely functional inactivation, to subtle or difficult-to-detect effects, or possibly to altering the location, stability or translation of a transcript, including oncogenes and tumor-suppressor genes. Alternative splicing has not only critical roles in normal development but also is indispensable in multiple pathological processes, including cancers (2–4). Previous studies have provided evidence that aberrant splicing patterns are closely related to tumor progression and prognosis (2). For example, alternative splicing in pre-mRNA of Epidermal Growth Factor Receptor (EGFR) produces several isoforms, some of which are constitutively active, leading to enhanced tumorigenicity, migration, and invasion (5, 6). EGFR, Insulin Receptor (INSR), and Vascular Endothelial Growth Factor Receptor (VEGFR), whose alternative splicing features variated, result in promoting tumor progression or reduced response to therapy (7). Recent evidence found that several tumor suppressor genes undergo aberrant AS in cancer, which leads to

either complete or partial loss of function, such as TP53 (8). Therefore, alternative splicing events might be ideal biomarkers for cancer diagnosis and prognosis and even be served as a potential target which might help scientists to discover new drugs.

The conventional molecular method for quantification of AS is a reverse transcription polymerase chain reaction (RT-PCR). There are several other techniques, including expressed sequence tags (ESTs) and splicing-sensitive microarrays, which were invented to identify the connections between genotypes and AS patterns in patients. However, these technologies have low throughput, high noise, or restrained to known splicing events. Powered by high-throughput RNA-seq, the amount of human transcriptome data has grown tremendously over the past decade, and large-scale studies in aberrant AS events at a more fine-grained level are now available. Recent advances in RNA-Seq and related bioinformatics methods allow researchers and clinicians to discover cancer-related AS and further investigate the molecular mechanism.

Pancreatic cancer is still known as one of the most malignant solid tumors whose 5-year survival rate has remained under 8% over the past 30 years. The disease is typically found at a late stage when the resection is impossible. Moreover, a response rate of only one-quarter or less can be expected, and resistance of current chemotherapy, such as gemcitabine, occurred in most of the pancreatic cancer patients. At present, the molecular mechanism of pancreatic cancer development and progression is still unclear. Researches have been undertaken to elucidate the mechanisms of this malignancy, including AS in specific gene transcription (9–11). However, few studies have tried to investigate the prognostic value of AS in pancreatic cancer. Therefore, the present study identified pancreatic cancer-specific AS events by analysis of RNA-seq data downloaded from The Cancer Genome Atlas (TCGA) program, gaining more information about their functions in cancer biology in detail.

MATERIALS AND METHODS

Alternative Splicing Events From TCGA RNA-Seq Data

TCGA (https://tcga-data.nci.nih.gov/tcga/) is a landmark cancer genomics program with a large amount of detailed information across various cancers in public database (12). The RNA-Seq data of pancreatic cancer cohorts (PAAD) was downloaded for further analysis. SpliceSeq (http://bioinformatics.mdanderson.org/TCGASpliceSeq) is a Java application which explores the mRNA alternative splicing patterns of TCGA data. The SpliceSeq tool was used to investigate the mRNA splicing pattern of PAAD samples from the TCGA database. SpliceSeq aligned reads to available transcripts of genes in the Ensembl database and built a unified splice graph. Then, the PAAD sample reads are aligned to the splice graph, and the feature of splicing for each transcript will be summarized. The Percent Spliced in (PSI) value is a parameter to assess the chance of each splicing event. There are several subtypes of splice events: Exon Skip (ES), Alternate Promoter (AP), Mutually Exclusive Exons (ME), Alternate Terminator (AT), Retained Intron (RI), Alternate Donor site (AD), and Alternate Acceptor site (AA). The detailed information of each subtype of splicing event in PAAD was shown in **Figure 1A**.

Survival Analysis

Clinical information of the PAAD cohort with 178 patients was available in the TCGA database (12). Summary characteristics of these patients were shown in **Supplementary Table 1**. In order to build the model and further analysis, we used mean values to replace the null value in the dataset of the splicing events. For each AS event, the patients were divided into two groups according to the median value; then the Univariate Cox analyses were performed to identify survival- associated splicing AS events in pancreatic cancer ($P < 0.05$). The Multivariate Cox regression was performed to determine the prognostic value of splicing events ($P < 0.05$). Then, the most significant top 20 genes in each model were chosen for the forest plots. Above analyses were performed using R/Bioconductor (version 3.5.2) and SPSS (version 25.0).

Construction of the Model of Risk Scores

Predictive models were built with prognostic events from identical AS subtype, respectively, whereas the final model was constructed with the whole splicing events from PAAD. In order to evaluate accuracy of model of risk scores, we drew the K-M curve, and the cut-off value is $P < 0.01$. Receiver-operator characteristic (ROC) curves were drawn, and the values of the area under the curves (AUC) were used to compare the predictive power of each model. All analyses were performed using R/Bioconductor (version 3.5.2) and Graphdpad Prism 8.0.

UpSet Plot and Gene Network Construction

Intersections between different types of AS were investigated by UpSet R (13). UpSet R is a novel R package which provides intersecting sets using matrix design, along with visualizations of several common sets, element, and attribute related tasks. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway were performed and were significant when the *P*-value was <0.05 in KEGG and 0.0001 in GO analysis. GO Enrichment plot were used to depicted gene interaction network, function annotation, and pathway enrichment of survival- associated AS genes. Therein, using Cytoscape (version 3.7.1), significant genes with the smallest *P*-value in univariate analysis were selected for the drawing of the PPI network.

Splicing Correlation Network Construction

The expression of splicing factor genes in mRNA splicing pathway was investigated by analysis of the level 3 mRNA-seq data in TCGA. Pearson correlation test was used to analyze the correlation between the mRNA expression of splicing factor gene and the PSI value of survival- associated alternative splicing events. Cytoscape (version 3.7.1) was used to construct the interaction network of the significant genes with the smallest *P*-value.

Analysis of Splicing-Factor, Transcription Factors, and Driver Gene

The association between survival- associated AS events and splicing factors was further investigated. Firstly, the log-rank test was used to identify survival- associated splicing factors. The list of 71 known splicing factors was extracted from the SpliceAid 2 (https://bioinformatics.mdanderson.org) database, which was released in February 2013 (14). The expression profiles of splicing factors were downloaded from the TCGA database and further converted into transcripts per million (TPM). Pearson correlation test was applied to assess the association between survival-associated AS and survival- associated splicing factors. FunRich (Functional Enrichment analysis tool for transcription factors) from ExoCarta (http://www.exocarta.org/), DriverDBv2 (A database for human cancer driver gene research) and David (http://david.abcc.ncifcrf.gov/) databases were used to perform the analysis. To find the correlation between gene mutation status and AS events, *t*-test was performed. Pearson correlation test was also performed to investigate the association between mRNA expression of driver genes and AS events. R software (version 3.5.2) was applied for bioinformatics analysis, and $P < 0.05$ was considered significant (Two-sided tests).

RESULTS

Number of mRNA Splicing Events in PADD Cohort From TCGA

The PSI value of all the splicing events was calculated by SpliceSeq. To identify each AS event precisely, each AS event was named by gene name followed by the unique as_ID and AS types. For example, for the name S100A13/7733/AP, S100A13 is the gene name, 7733 is the as_ID in the dataset, and the AP is the AS subtype. As depicted in **Figure 1B**, a total of 10,623 genes with 45,313 AS events were detected in 178 pancreatic samples, including 17,402 ESs in 6,750 genes, 2,873 RIs in 1,922 genes, 9,325 APs in 3,724 genes, 8,733 ATs in 3,816 genes, 3,118 ADs in 2,210 genes, 3,657 AAs in 2,594 genes, and 205 MEs in 202 genes. Overall results showed that one gene might have an average of 4.2 AS events. Among those genes, 8,833 genes had more than one type of AS events. Gene collagen type 1 alpha 1 (COL1A1) had the maximum number of AS events ($n = 484$), followed by mitochondrial ribosomal protein L55 (MRPL55) ($n = 74$) and interleukin 32 (IL32) ($n = 68$). Among those splicing subtypes, ES was the main subtype of AS events, while ME was relatively rare in the tumor. Besides, only a small proportion of AS events (1,622 out of 45,313) were novel splice. The PADD cohort of TCGA also included four normal samples; the PSI median values of different genes were also summarized and further analyzed. Several genes splicing events, including KIAA1715/56096/AP, ZNF567/49415/AP, NTMT1/87861/AP, ANAPC15/17570/AD, SRPK2/81284/ES, MTMR11/7413/AP, FNIP2/70999/AP, and TNC/87336/ES, differed significantly between tumor and normal samples (**Figure 2A**). When compared to normal samples, cancer samples had reduced alternative splicing diversity (41,629 AS events in normal vs. 40,959 in cancers).

Survival-Associated AS Events in PAAD Cohort

Cox univariate analyses of overall survival were applied to explore survival- associated AS events in PAAD cohort. The results showed that 6,711 AS events strongly correlated with OS ($P < 0.05$), including 550 RIs from 449 genes, 421 AAs from 382 genes, 385 ADs from 342 genes, 1,499 APs from 809 genes, 1,649 ATs from 873 genes, 2,174 ESs from 1,463 genes, 33 MEs from 33 genes and 550 RIs from 449 genes. The UpSet plot was a novel method to display the intersecting sets, which may be more intuitive and superior to the Venn diagrams. As depicted in the plot, most of these genes had two or more AS subtypes associated with survival, but none of them possessed seven AS subtypes simultaneously (**Figure 2B**). The top 20 survival-associated AS events of the seven AS subtypes were presented in **Figure 3**. In top 300 genes from survival-associated AS events, some genes were top hub genes in the network, such

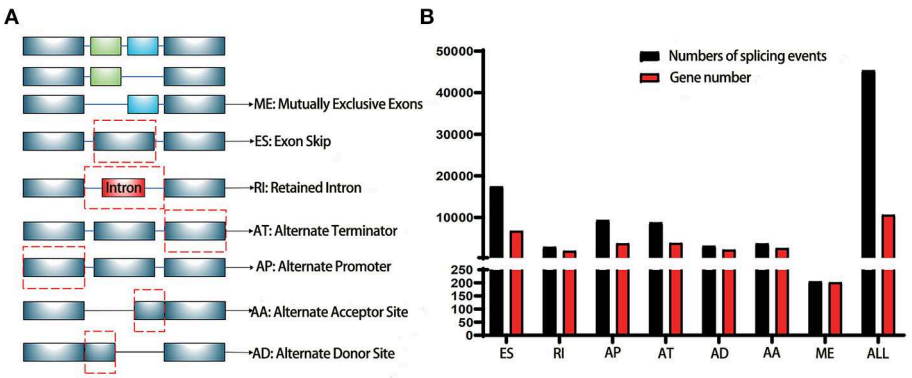


FIGURE 1 | Illustrations for alternative splicing during seven types in this study. **(A)** Schematic example of AS events, ME, Mutually exclusive exons; ES, Exon skip; RI, Retained intron; AT, Alternate terminator; AP, Alternate promoter; AA, Alternate acceptor site; AD, Alternate Donor site; **(B)** A number of AS events and involved genes from TCGA PAAD cohort were depicted according to the AS types. The black bar represents the preliminarily detected AS events. The red bar represents the related genes.

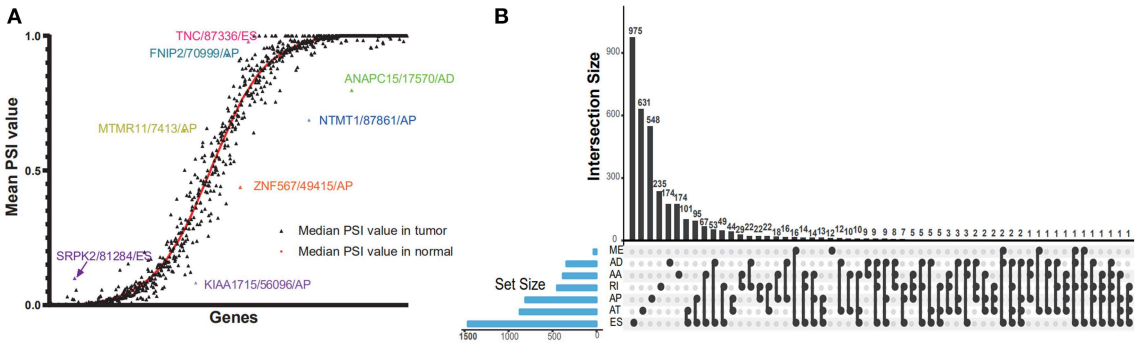


FIGURE 2 | Dot plot and UpSet plots in PAAD. **(A)** Correlation between tumor PSI and normal PSI in splicing factors were depicted in the dot plot. The smooth red curve was drawn according to median PSI value in normal; the black triangles represented the median PSI value of genes in the tumor. **(B)** The UpSet intersection diagram shows seven subtypes of splicing associated AS events in PAAD. One gene might have more than one subtype of survival-associated AS event.

as VEGFA, CD44, pyruvate kinase gene (PKM), amyloid beta precursor protein (APP), ubiquitin-conjugating enzyme E2 L6 (UBE2L6) (**Figure 4A**). In pancreatic cancer, KEGG pathway analysis showed that "Metabolic pathway," "Endocytosis," and "Axon guidance" were most significantly enriched by these genes. GO analysis revealed that "Protein binding," "poly(A) RNA binding," and "RNA binding" in molecular function, "cytoplasm," "cytosol," and "extracellular exosome" in cellular component, "cell-cell adhesion," "mRNA processing," and "actin cytoskeleton organization" in biological process were the most significantly enriched (**Figure 4B**).

Prognostic Models for PADD Cohort

To evaluate the prognostic value of AS events in pancreatic cancer, the survival-associated AS events were selected to construct the prognostic risk score models in each subtype of AS events (**Figure 5**). As depicted in the results, all of the models showed significant value to predict the outcome of pancreatic cancer patients, including RI subtype ($P < 0.0001$), ES subtype ($P < 0.0001$), AP subtype ($P < 0.0001$), AT subtype ($P < 0.0001$), AA subtype ($P < 0.0001$), ME subtype ($P < 0.0001$), and AD subtype ($P < 0.0001$) (**Figure 6A**). The final prognostic model was built by a combination of prognostic AS events from different subtypes and showed significant prognostic value in distinguishing high-risk patients ($P < 0.0001$). Notably, the final prognostic model showed better performance than seven AS subtypes. The final prognostic predictor had the highest predicting efficiency analyzed by ROC (AUC = 0.89), followed by the AP model in subtypes (AUC = 0.88) (**Figure 6B**).

Network of Survival-Associated Splicing Factor, Transcription Factors, and Driver Gene

To identify survival-associated splicing factors, we performed a survival analysis about splicing factors based on PSI values. A total of 71 splicing factors from the SpliceAid2 database were chosen for survival analysis. Results showed that AS events of five splicing factors, including DAZ associated protein 1 (DAZAP1), RNA-binding motif 4 (RBM4), Epithelial Splicing Regulatory Proteins 1 (ESRP1), Quaking (QKI), and steroidogenic factor 1 (SF1), significantly associated with overall survival. The level 3 RNA sequence data were downloaded from TCGA, and the correlations of splicing factors expression and survival were analyzed. As depicted in **Figures 7A–E**, the expression of ESRP1 ($P = 0.0025$) significantly associated with survival, but DAZAP1 ($P = 0.064$), QKI ($P = 0.45$) and SF1 ($P = 0.62$) and RBM4 ($P = 0.18$) were not. The association between PSI values of top significant AS events and survival-related splicing factors was still unknown. Thus, String tool was used to investigate the association and gain systematic insights into their interaction. Only genes that are significantly related to each other were included in the network. In the correlation network, there was a significant association between the expression of five survival-associated splicing factors and 95 survival-associated AS events. Among 95 survival-associated AS events, 56 AS events (green dots) predicted good survival, whereas 39 AS (red dots) events strongly associated with poor survival in pancreatic cancer (**Figure 7F**). Correlation between these five splicing factors and representative AS events was shown in dot plots, suggesting the potential association between them (**Supplemental Figure 1**).

A transcription factor enrichment prediction performed among the survival-associated AS events using the FunRich software. Results identified several transcription factors, including Krüppel-like factor 7 (KLF7), GA binding protein transcription factor subunit alpha (GABPA), trans-acting transcription factor 1 (SP1), that might be the most significant transcription factors associated with survival-associated AS events. Transcription factor SP1 was the most highly related to 53.4% of all the survival- associated AS genes, followed by KLF7 (36.5%) and GABPA (23.9%) (**Figure 8A**).

A list of driver genes was generated by at least five bioinformatics tools using the DriverDB, which is a database for the investigation of cancer driver gene and mutations. Results showed that 13 driver genes were identified, including tumor protein p53 (TP53), which were previously reported (15) (**Figure 8B**). In the mutation profile of driver genes, mutation of TP53, FSHD region gene 1 family member B (FRG1B), and cell division cycle 27 (CDC27) occurred in most of PAAD

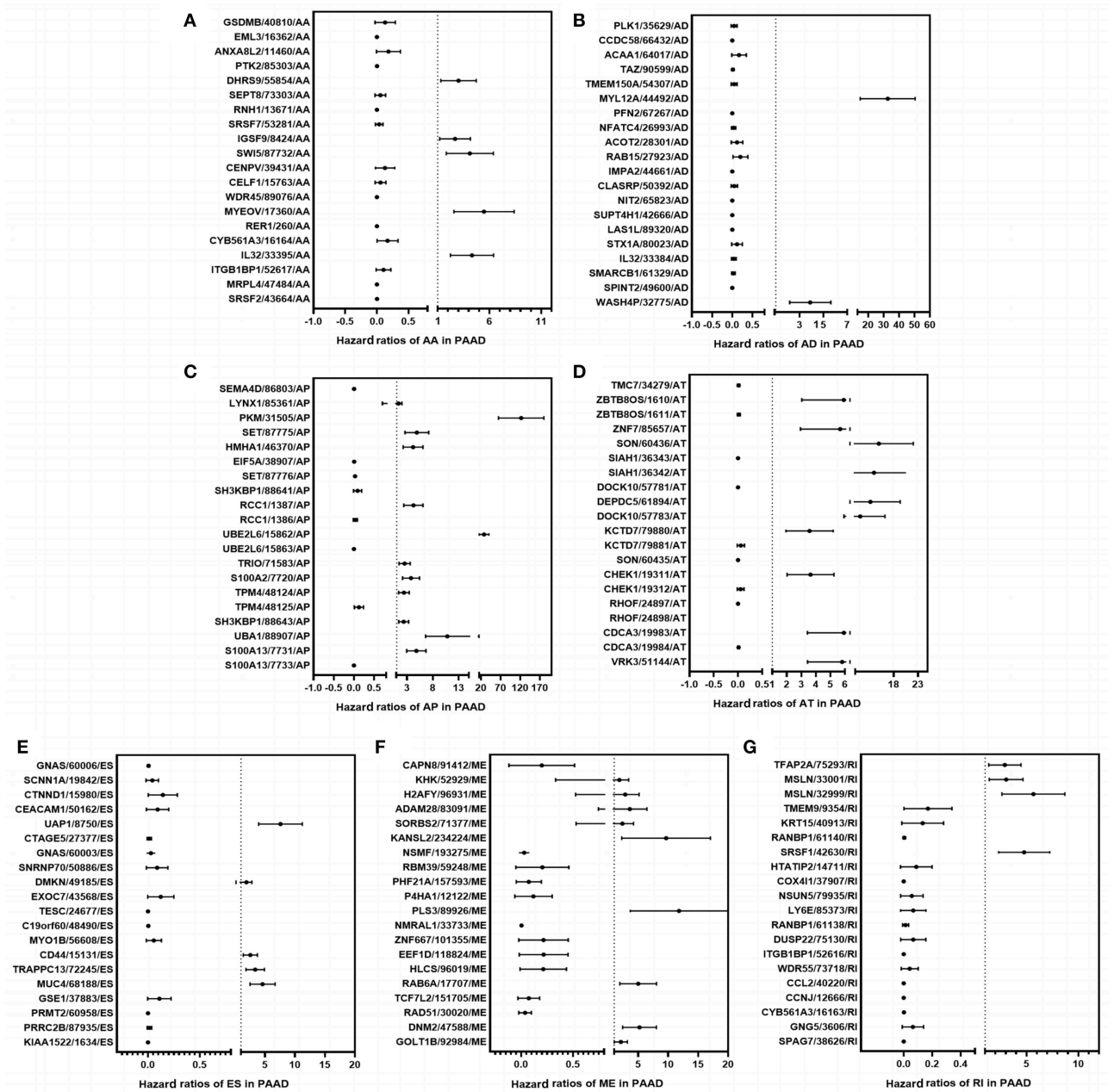


FIGURE 3 | Forest plots show the top 20 survival-associated AS events of the seven AS subtypes, respectively. The circles represent HRs in the plots; Horizontal bars represent 95% CIs. Forest plots of HRs for survival associated AA subtypes **(A)**, AD subtypes **(B)**, AP subtypes **(C)**, AT subtypes **(D)**, ES subtypes **(E)**, ME subtypes **(F)**, RI subtypes **(G)** in PAAD.

cohort from TCGA. As for the mutation class, truncating and missense were the two main types for driver genes, such as TP53, FRG1B, and CDC27 (**Supplemental Figure 2**). In addition, we investigated the correlations between mRNA expression of driver genes and the top 30 survival-associated AS events. Results indicated that mRNA expression of adaptor-related protein complex 3 subunit sigma 1 (AP3S1), integrin subunit beta 4 (ITGB4), and p21 (RAC1) activated kinase 1 (PAK1) was significantly associated with most of the top 30 survival-associated AS events (**Supplemental Figure 3**). Samples were divided into several groups according to numbers of driver gene mutations, and results indicated that numbers of AS events for each sample were not significantly associated with numbers of driver gene mutations (**Supplemental Figure 4**). Furthermore, we explored the correlation of AS events and mutation profiles by the *t*-test and found that mutation status of TP53, splicing factor

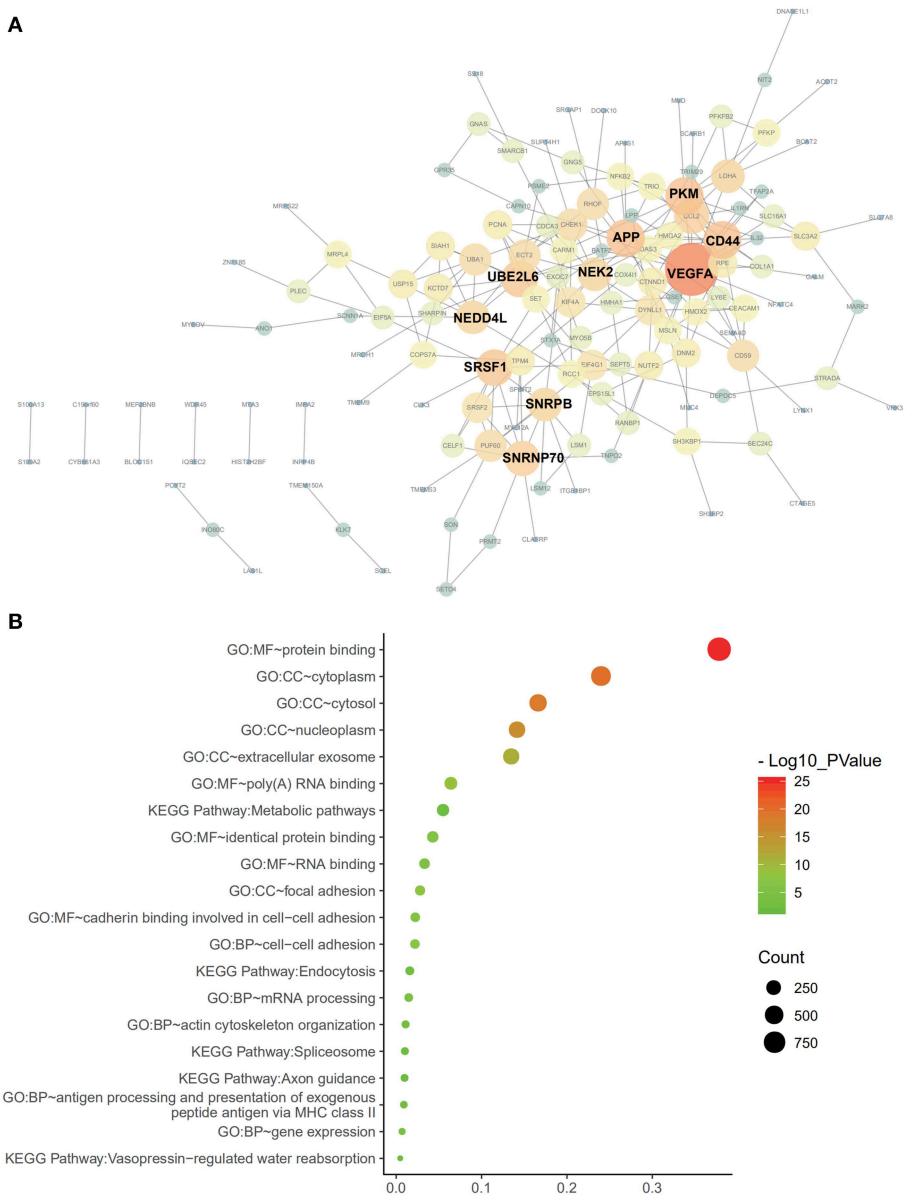


FIGURE 4 | Protein-protein interaction analysis and gene enrichment in PAAD. **(A)** Survival-associated AS events interaction network created by Cytoscape. Genes are represented as nodes in the plot, and their interactions were denoted by lines. The size and color of the nodes represent Degree values and change pattern, respectively. The gene of lighter color and greater circle shows the higher Degree values in this network, whereas the darker color and the smaller circle show the smaller Degree values in this network. **(B)** Pathways identified by GO and KEGG analyses. Top 15 enrichment analysis of GO (include BP, CC, and MF, respectively) and top five pathways KEGG analyses of genes from OS-related alternative splicing events. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; CC, cellular component; Mf, molecular function; BP. Biological process.

3a subunit 1 (SF3A1), and CDC27 significantly correlated with most of the Top-100 survival-associated AS events (**Figure 9**).

DISCUSSION

Alternative splicing enables a single gene to generate multiple mRNAs. Moreover, these mRNAs can be translated into various proteins with diverse functions and structures. Emerging data have demonstrated that aberrant AS patterns were identified in various cancers and engaged in multiple carcinogenic processes during cancer development and progression (16). The previous study demonstrated the AS events of tissue factors promoted neovascularization and monocyte recruitment via integrin ligation, thus contributing to activation of coagulation and tumor spread in pancreatic cancer (17). In pancreatic cancer, AS events of the PKM were differentially regulated and promoted the expression of the PKM2 isoform. Compared to PKM1, switching PKM2 AS events is beneficial to withstand

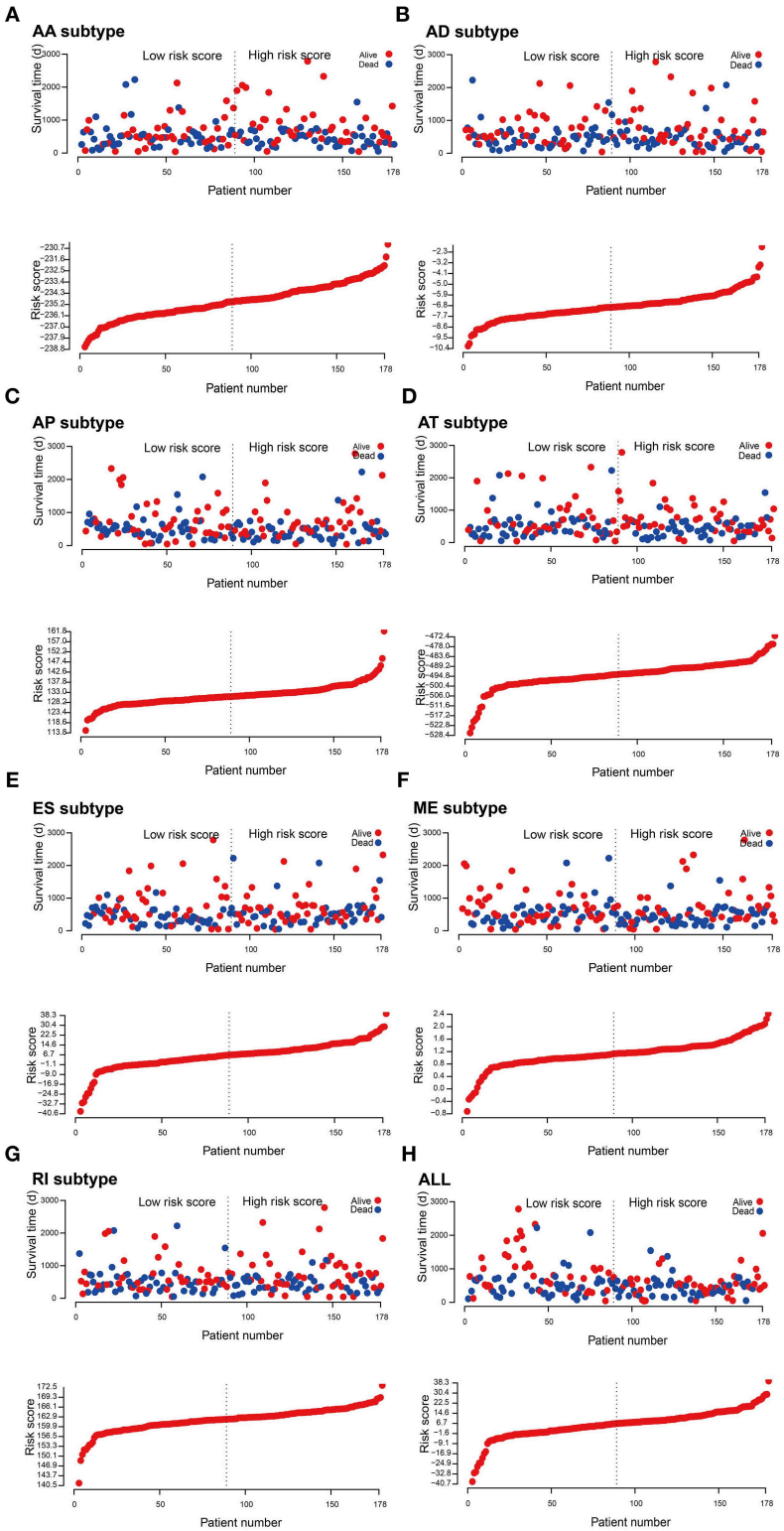


FIGURE 5 | Construction and analysis of risk score based on the survival-associated splicing events using multiple Cox regression analysis. PAAD patients were divided into low- and high-risk groups based on the median value of risk score. The top of each assembly drawing represents survival status and survival time of PAAD patients distributed by risk score, the bottom part is the risk score curve of patients with PAAD. Risk scores were constructed using **(A)** AA subtypes, **(B)** AD subtypes, **(C)** AP subtypes, **(D)** AT subtypes, **(E)** ES subtypes, **(F)** ME subtypes, **(G)** RI subtypes, and **(H)** ALL subtypes of survival-associated splicing events.

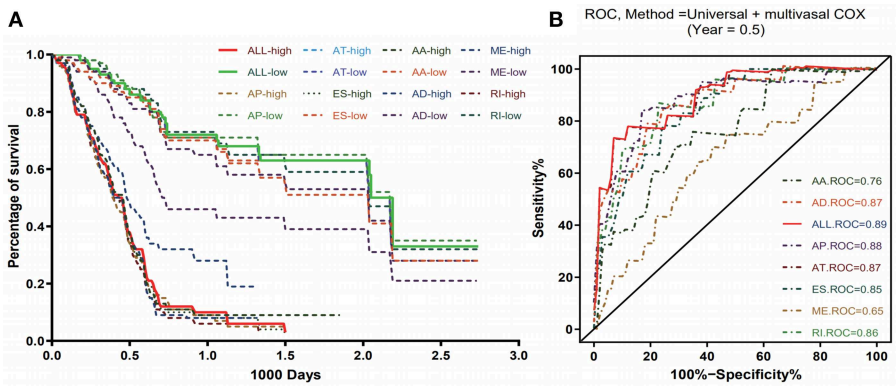


FIGURE 6 | Kaplan-Meier and ROC curves of prognostic predictors in PAAD cohort. **(A)** Kaplan-Meier plot depicting the survival difference between the high and low-risk group in these prognostic models. **(B)** ROC analysis for all prognostic models. The different color lines of ROC curves represent different subtypes of AS events.

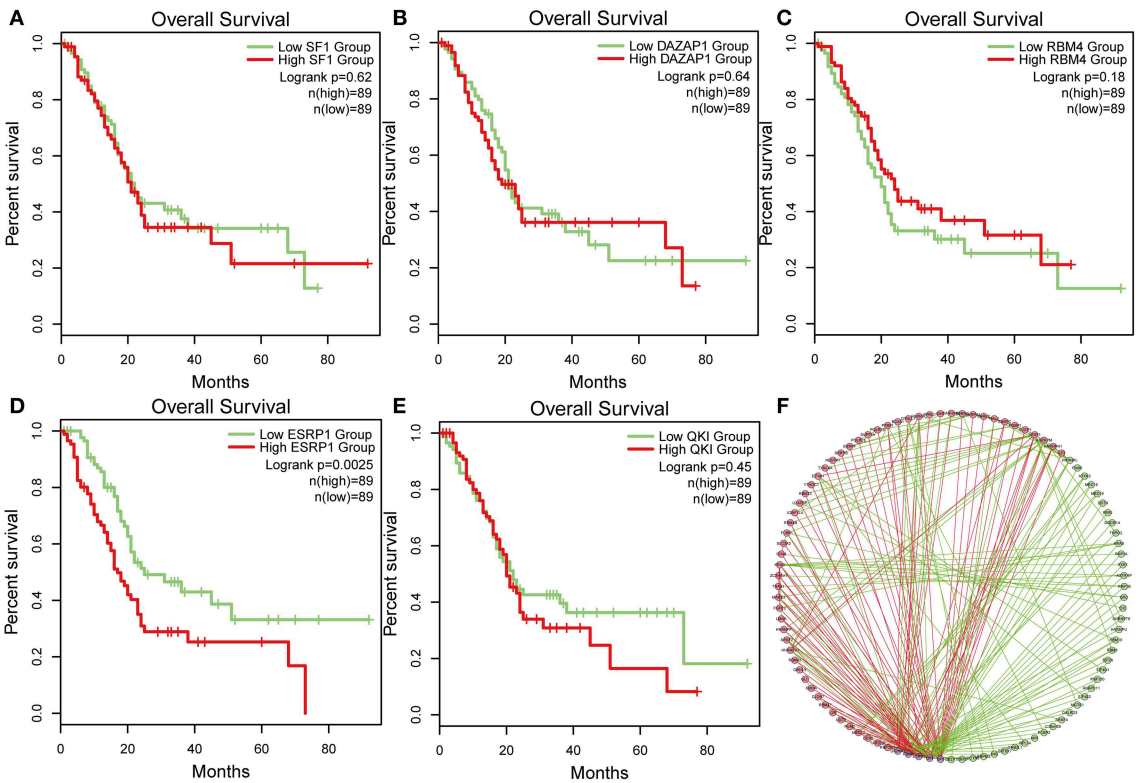


FIGURE 7 | Survival-associated splicing factors and splicing correlation network in PAAD. **(A–E)** The prognostic value of mRNA expression of five splicing factors expression, whose AS events was significantly associated with overall survival in PAAD. **(F)** Splicing correlation network in patients with PAAD constructed by Cytoscape. These five splicing factors (purple dots) were positively (red lines) or negatively (green lines) associated with AS events, which predicted good (green dots) or poor (red dots) outcomes in patients with PAAD.

gemcitabine and cisplatin-induced genotoxic stress, thus induced chemoresistance (18). Serine and arginine-rich splicing factor 1 (SRSF1) and heterogeneous nuclear ribonucleoprotein K (hnRNPK) were aberrantly upregulated in pancreatic cancer, leading to the increased expression of anti-apoptotic splice variants of Bcl-x and Mcl-1, significantly affected responses to chemotherapy (19). Previous data concerning the function of AS events in pancreatic cancer mainly focused on one or several genes, and there was no study which had explored the prognostic value of AS comprehensively. Given the importance of AS events in cancer, we investigated AS events and gained a comprehensive insight into the prognostic value

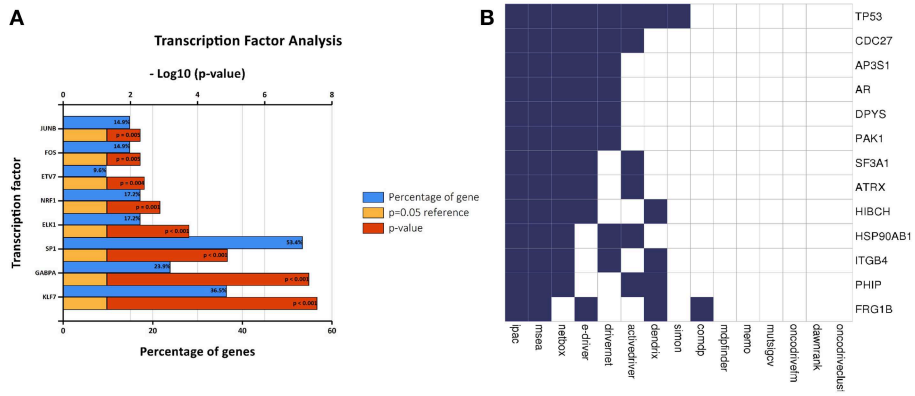


FIGURE 8 | Correlation between transcription factors, driver mutation and splicing factors. **(A)** The histogram shows the results of transcription factor prediction from survival- associated AS events. The blue band represents the gene percentage, the yellow band represents the *P*-value standard ($P = 0.05$), and the red band represents the *P*-value. **(B)** A list of driver genes was generated by at least five bioinformatics tools using the DriverDB.

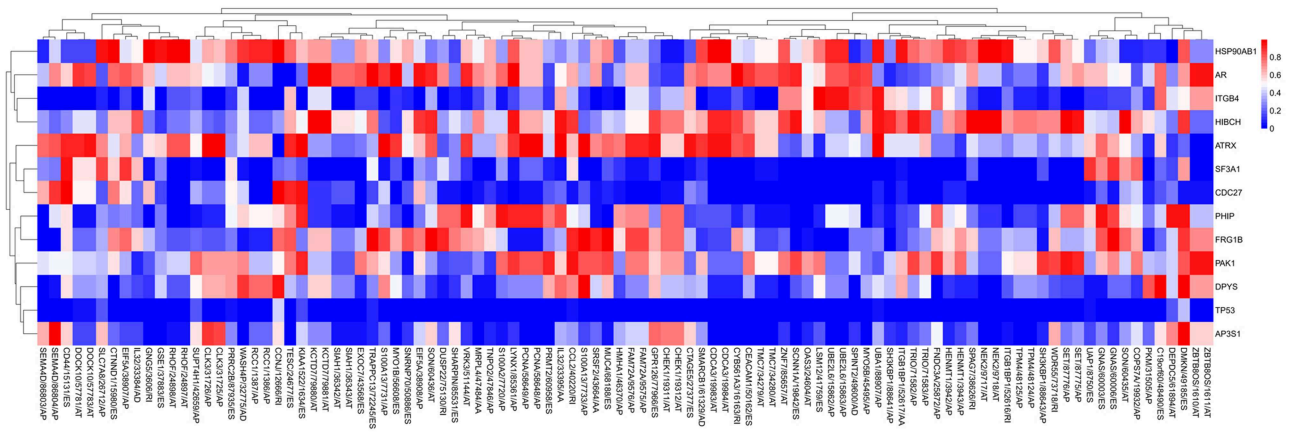


FIGURE 9 | The correlation between PSI value of AS events and mutated status of driver genes was explored through *t*-test. Colors represented the *P*-value of *t*-test. Blue to red means *P*-value from low to high.

of AS events in pancreatic cancer through the analysis of TCGA.

Among the genes with AS events, Gene COL1A1, which makes part of a large molecule called type I collagen, have the maximum number of AS events. Further analysis revealed some of COL1A1 AS events significantly correlated with survival. Our results were consistent with previous studies (20–22). Evidence showed that COL1A1 could activate β1-integrin and the activation, along with the epithelial-mesenchymal transition, contributed to the development of PAAD (23). The previous study has also demonstrated that once PAAD cells met COL1A1, Snail expression conducted by the increasing of TGF-β1 (Transforming Growth Factor-β1) signaling would begin, which in turn accelerate the progress of PAAD invasion by the upregulated MT1-MMP (membrane type 1-MMP) expression (24). Evidence also showed that hypoxia augmented the transcription and deposition of COL1A1 by TGF- β pathway, and COL1A1 was identified as a hypoxia marker in the non-small cell lung carcinoma (20). Abnormal COL1A1 lead to increasing radioresistance in cervical cancer and had its potential prognostic value in gastric cancer (21, 22). However, the implication of dysregulated splicing pattern of COL1A1 in cancer, including pancreatic cancer with abundance fibrosis, remains to be elucidated. When compared to normal samples, cancer samples had reduced alternative splicing diversity. A previous study reported that the splicing factor genes were upregulated in seven cancer types, including colorectal adenocarcinoma, breast cancer, and lung adenocarcinoma, while they were downregulated in four cancer types, including lymphoma and uterine cancer (2). In our study, we found that the total expression of the splicing factor genes in pancreatic cancer was downregulated. The results indicated that dysregulated expression of the splicing factor genes among cancer types was not in a fixed mode, which may partly result from tumor heterogeneity. Thus, systemic evaluation of the AS patterns in pancreatic cancer contributes

to the understanding of the underlying mechanism of tumor development and progression.

Survival analysis was conducted, and interaction analysis between these survival-associated genes was performed. Results indicated that VEGFA closely related to other genes and served as a hub gene in the network. Among the VEGFA AS events, patients with VEGFA/76330/ES had better survival, implying that loss of Exon8 may weaken or abolish the interaction of VEGFA with other proteins and then inhibit the growth of the tumor. However, VEGFA/76336/ES significantly associated poor survival in pancreatic cancer, which is inconsistent with previous data (25). Of note, VEGFA/76336/ES, whose splice occurred with removal of exon7.1 and exon7.2 loss, lack the neuropilin binding site at exon7. In breast cancer, the VEGF-A/Neuropilin 1 pathway promoted cancer stemness by activating Wnt/β-Catenin axis, resulting in cancer stem cell phenotypes and chemoresistance (25). In acute myeloid leukemia, high expression of VEGFA was identified as an oncogenic factor, whose function may be reversed by SEMA3A competing for neuropilin (26). Theoretical speaking, removal of exon7, the binding site of neuropilin at VEGF sequence, abolish the interaction and inhibit tumor growth. However, VEGFA/76336/ES significantly associated unfavorable prognosis, which indicating its multifaceted roles in pancreatic cancer progression. It is hard to conclude that VEGFA/76336/ES promotes tumor growth due to a lack of experimental evidence. Nevertheless, our results indicated that neuropilin mediates cancer cell growth may rely on pathways independent of VEGFA. Additionally, blocking neuropilin may strengthen the role of anti-VEGF therapy in reducing the formation of new blood vessels. It is difficult to judge whether a gene is a cancer suppressor or a promoter since different AS events have varied, even opposite biological functions. Therefore, mRNA expression of a gene may be not adequate to determine the biological function, and the predominant AS events need to be taken into account.

Due to the characteristics of pancreatic cancer, including late diagnosis and poor outcome, several researchers had proposed some prognostic models based on mRNA, lncRNA, and microRNA (4, 27, 28). Nevertheless, seldom of these prognostic models come into widely used in clinical practices. Several studies published before finding that alternatively spliced variants contributed to cancer metastasis, cell cycle progression, and chemoresistance (18, 29, 30). As events have been previously identified as diagnostic, predictive, and prognostic biomarkers in pancreatic cancer (18, 31, 32). However, current knowledge about AS events was mostly derived from small samples studies or mainly focused on one single gene. Recently, a systemic analysis of AS events in pancreatic cancer was available due to high-throughput sequencing analysis and data from TCGA. Analysis of each subtype of splicing events was performed and found some of the AS events were of significant prognostic value in pancreatic cancer. Unlike other cancers, including colorectal cancer, lung cancer, the majority of AS events were closely associated with favorable prognosis in pancreatic cancer, especially in AD and RI subtypes. Prediction models were further built by each subtype, respectively or a combination of these seven subtypes. Among the models built by identical subtype, AP events demonstrated the highest efficiency in the prediction of survival outcome than other six subtypes. Moreover, the final prediction model built by a combination of seven subtypes showed better performance than other prediction models, with an AUC of ROC reaching 0.89 in distinguishing poor survival outcome. Our current work is the first to provide a comprehensive and systemic analysis of AS events and risk score models based on survival-associated AS events in pancreatic cancer.

The network of survival-associated splicing factors was evaluated and found AS events of DAZAP1, RBM4, ESRP1, QKI, and SF1 were significantly associated with overall survival, but the only mRNA expression of ESRP1 correlated with overall survival. Therefore, investigation into the AS events is important to judge the function of gene products. Epithelial-mesenchymal transition (EMT) is defined as a process that epithelial cells with tight junctions acquire a mesenchymal phenotype (33). This means that epithelial cells become easily mobile after this transition, that is, EMT can regulate metastasis (34). ESRP1 is a critical regulator in the epithelial splicing program through targeting several genes, such as fibroblast growth factor receptor 2 (FGFR2) and CD44 (also called H-CAM) (35, 36). As the levels of the mRNA of ESRP1 is down-regulated, the CD44 variant isoform is replaced by the CD44 standard isoform which promotes EMT, increasing invasiveness in gallbladder cancer (37). Evidence showed that the role of inflammation-inducible Snail in the driving malignant transformation of both normal and at-risk human bronchial epithelial cells required the silencing of RNA splice regulator ESRP1 (38). However, the evidence about the function of ESRP1 in pancreatic cancer still lacks and further studies are required. Current evidence has pointed out that splicing factors can precisely bind to a splice-regulatory sequence located at the gene, thus control the process of splicing (39). According to the difference in the sequence and structure, these splicing factors can be divided into two families, including Ser/Arg rich proteins (SR proteins) and the heterogeneous nuclear ribonucleoproteins (hnRNPs). By binding to sequence silencers or enhancers of splicing, these two families possess the opposite function in the mRNA splicing. However, the potential regulatory network of splicing factors during the splicing process remains unclear and clarifying the function of ESRP1 is critical in the interpretation of the molecular mechanism of pancreatic cancer. More attention should thus be paid to the study of AS events in pancreatic cancer.

The transcription process can impact AS events by a variety of mechanisms. Transcription factors can regulate the recruitment of splicing components, and modulate Pol II elongation rate, which regulates the kinetics of exposure of competing for splice sites (40). We evaluated the association between survival-associated AS events and transcription factors. Transcription factors KLF7, GABPA, and SP1, were the most highly related to survival-associated AS genes, which implied that one transcriptions factor might participate in splicing control of several genes. Krüppel-like factors (KLFs) was involved with many cellular activities, such as proliferation and metabolism (41–43). Moreover, a previous study reported

that KLF7 transcriptionally activated argininosuccinate lyase, which resulted in polyamines production and the oncogenesis of glioma (44). KLF7 can also contribute to the migration and epithelial-mesenchymal transition of oral squamous cell carcinoma (45). However, the mechanism of how transcriptions factors engaged in the process of splicing is still unknown. It is reasonable that one single transcriptions factor may regulate several genes not only by direct binding to the promoter of targeted genes but also by indirect impact on splicing process.

Recent evidence showed that several genetic mutations, including K-Ras, TP53, SMAD family member 4 (SMAD4), and cyclin dependent kinase inhibitor 2A/P16 (CDKN2A/P16), drove the oncogenesis of pancreatic cancer (46). Except for these four driver genes, more and more genes are identified as the critical genes in the process of pancreatic cancer, including ret proto-oncogene (RET), AT-rich interaction domain 1A (ARID1A), and ATM (47). Driver genes have been identified as the building blocks in pancreatic cancer, and emerging data suggested that driver gene K-Ras involved in the process of splicing control, such as mucin 6 (MUC6), hepatocyte growth factor (HGF), VEGFR-2, and VEGFB (48). The abnormal expression of splicing factors of SR and hnRNP families results in dysfunction of targeting apoptotic genes, including p53 (19). However, rare studies had been conducted in the exploration of the association between driver genes and AS events. Potential driver genes were identified by the bioinformatic tool in the present study. Further analysis revealed that splicing events of each gene did not increase with accumulating gene mutations. Though the expression of TP53 and SF3A1 correlated with rare survival-associated AS events, mutation status of these two driver genes significantly correlated with many of the top 100 survival-associated AS events. SF3A1, which belong to candidate U2-dependent spliceosome genes family, was identified as driver genes by five prediction tools. Previous studies indicated that two SNPs (rs5994293 and rs9608886) of SF3A1, locating to the region of 22q12.2, were strongly correlated with pancreatic cancer (49). However, the mechanism of how driver genes, including SF3A1, lead to increasing AS events is still unclear. Our study findings enriched our knowledge about the mutation status of driver genes and regulation of splicing, gaining systemic insight into the molecular mechanism underlying PAAD.

Several limitations should be considered when interpreting the results. First, the included number of the PAAD samples was relatively small, and only four normal samples were available for PSI analysis. Second, the prognostic value of survival-associated AS events lack the external independent validation cohort. Third, the present study only investigated the data from high-throughput genomic sequence; experimental validation should be performed in the future.

In conclusion, our comprehensive investigation first focused on the aberrant AS patterns in pancreatic cancer and may contribute to the improvement of pancreatic cancer management and broaden to the novel field of prognosis and targeted molecular implications.

DATA AVAILABILITY

The original data of the present study can be found at TCGA (https://tcga-data.nci.nih.gov/tcga/) and the SpliceAid 2 (https://bioinformatics.mdanderson.org).

AUTHOR CONTRIBUTIONS

MY contributed to conception and design, and acquisition, analysis, and interpretation of data. WH contributed to the acquisition of data of acquisition and data analysis. SR contributed to the acquisition of data, analysis, and interpretation of data. RG has been involved in drafting the manuscript and revising it critically. LT contributed significantly to drafting the manuscript. BHu contributed to acquisition of data. BHo contributed to revising the manuscript. ZJ contributed to interpretation of data. LM contributed to data interpretation. HJ conducted the study. All the authors participated in the discussion and editing of the manuscript.

FUNDING

This grant of the study was from the National Natural Science Foundation of China (Grant No. 81701560), National Science Foundation of Guangdong Province, People's Republic of China (Grant No. 2017A030313530) and Guangzhou Science and Technology Plan of Scientific Research Projects, People's Republic of China (Grant No. 201904010021). These fundings made a significant contribution to study design, data interpretation, and writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.00773/full#supplementary-material

Supplemental Figure 1 | Correlation between these five survival-associated splicing factors and representative AS events was shown in dot plots.

Supplemental Figure 2 | The mutation profile of 13 driver genes. The red band represents truncating, the purple band represents missense, and the green band represents inframe.

Supplemental Figure 3 | The heatmap of the correlations between the mRNA expression of driver genes and PSI values of top 30 survival-associated AS events. Colors represented the correlation coefficient *r*.

Supplemental Figure 4 | Samples from PAAD cohort were divided into several groups according to numbers of driver gene mutations from 0 to 12 in X-axis. No sample has thirteen gene mutations concurrently. The Y-axis represents the numbers of AS events of each sample.

Supplementary Table 1 | Baseline characteristics according to TCGA Clinical data.

REFERENCES

1. Wang ET, Rickard S, Shujun L, Irina K, Lu Z, Christine M, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. (2008) 456:470–6. doi: 10.1038/nature07509
2. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*. (2015) 35:2413–27. doi: 10.1038/onc.2015.318
3. Aversa R, Sorrentino A, Esposito R, Ambrosio MR, Amato A, Zambelli A, et al. Alternative splicing in adhesion- and motility-related genes in breast cancer. *Int J Mol Sci*. (2016) 17:121. doi: 10.3390/ijms17010121
4. Shi G, Zhang J, Lu Z, Liu D, Yang W, Wu P, et al. A novel messenger RNA signature as a prognostic biomarker for predicting relapse in pancreatic ductal adenocarcinoma. *Oncotarget*. (2017) 8:110849–60. doi: 10.18632/oncotarget.22861
5. Angélique G, Karine D, Hélène RC, Isabelle P, Laura M, Sandrine R, et al. Adult diffuse gliomas produce mRNA transcripts encoding EGFR isoforms lacking a tyrosine kinase domain. *Int J Oncol*. (2012) 40:1142–52. doi: 10.3892/ijo.2011.1287
6. Padfield E, Ellis HP, Kurian KM. Current therapeutic advances targeting EGFR and EGFRvIII in glioblastoma. *Front Oncol*. (2015) 5:5. doi: 10.3389/fonc.2015.00005
7. Abou Faycal C, Hatat A, Gazzeri S, Eymin B. Splice variants of the RTK family: their role in tumour progression and response to targeted therapy. *J Mol Sci*. (2017) 18:383. doi: 10.3390/ijms18020383
8. Surget S, Khoury MP, Bourdon J. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *OncoTargets Therap*. (2013) 7:57–68. doi: 10.2147/OTT.S53876
9. Chakedis J, French R, Babicky M, JnqfAQuish D, Howard H, Mose E, et al. A novel protein isoform of the RON tyrosine kinase receptor transforms human pancreatic duct epithelial cells. *Oncogene*. (2016) 35:3249–59. doi: 10.1038/onc.2015.384
10. Zheng KL, He TL, Ji WP, Jiang H, Shen Y, Li G, et al. Alternative splicing of NUMB, APP and VEGFA as the features of pancreatic ductal carcinoma. *Int J Clin Exp Pathol*. (2015) 8:6181–91.
11. Jiang P, Li Z, Tian F, Li X, Yang J. Fyn/heterogeneous nuclear ribonucleoprotein E1 signaling regulates pancreatic cancer metastasis by affecting the alternative splicing of integrin β1. *Int J Oncol*. (2017) 51:169–83. doi: 10.3892/ijo.2017.4018
12. Katarzyna T, Patrycja C, Maciej W. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. (2015) 19:68–77. doi: 10.5114/wo.2014.47136
13. Alexander L, Nils G, Hendrik S, Romain V, Hanspeter P. UpSet: visualization of intersecting sets. *Vis Comput Grap IEEE Trans*. (2014) 20:1983–92. doi: 10.1109/TVCG.2014.2346248
14. Francesco P, Matteo G, Alessandra Ballone B, Giovanni P. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum Mutat*. (2015) 33:81–5. doi: 10.1002/humu.21609
15. Minoru O, Keiichi O, Shinobu M, Reiji H, Takashi M, Yasuyuki S, et al. Immunohistochemically detected expression of 3 major genes (CDKN2A/p16, TP53, and SMAD4/DPC4) strongly predicts survival in patients with resectable pancreatic cancer. *Ann Surg*. (2013) 258:336–46. doi: 10.1097/SLA.0b013e3182827a65
16. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*. (2016) 17:257. doi: 10.1038/nrg.2016.10
17. Dusten U, Kevin T, Ramprasad S, Begüm K, Xiaoyang Q, Zhengtao C, et al. Alternatively spliced tissue factor contributes to tumor spread and activation of coagulation in pancreatic ductal adenocarcinoma. *Int J Cancer*. (2014) 134:9. doi: 10.1002/ijc.28327
18. Calabretta S, Bielli P, Passacantilli I, Pilozzi E, Fendrich V, Capurso G, et al. Modulation of PKM alternative splicing by PTBP1 promotes gemcitabine resistance in pancreatic cancer cells. *Oncogene*. (2016) 35:2031–9. doi: 10.1038/onc.2015.270
19. Kedzierska H, Piekiełko-Witkowska A. Splicing factors of SR and hnRNP families as regulators of apoptosis in cancer. *Cancer Lett*. (2017) 396:53. doi: 10.1016/j.canlet.2017.03.013
20. Oleksiewicz U, Liloglou T, Tasopoulou K, Daskoulidou N, Gosney JR, Field JK, et al. COL1A1, PRPF40A, and UCP2 correlate with hypoxia markers in non-small cell lung cancer. *J Cancer Res Clin*. (2017) 143:1133–41. doi: 10.1007/s00432-017-2381-y
21. Li J, Ding Y, Li A. Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer. *World J Surg Oncol*. (2016) 14:297. doi: 10.1186/s12957-016-1056-5
22. Liu S, Liao G, Li G. Regulatory effects of COL1A1 on apoptosis induced by radiation in cervical cancer cells. *Cancer Cell Int*. (2017) 17:73. doi: 10.1186/s12935-017-0443-5
23. Duan W, Ma J, Ma Q, Xu Q, Lei J, Han L, et al. The activation of β1-integrin by type I collagen coupling with the hedgehog pathway promotes the epithelial-mesenchymal transition in pancreatic cancer. *Curr Cancer Drug Target*s. (2014) 14:446–57. doi: 10.1158/1538-7445.PANCA2014-A39
24. Shields MA, Dangi-Garimella S, Krantz SB, Bentrem DJ, Munshi HG. Pancreatic cancer cells respond to type I collagen by inducing snail expression to promote membrane type 1 matrix metalloproteinase-dependent collagen invasion. *J Biol Chem*. (2011) 286:10495–504. doi: 10.1074/jbc.M110.195628
25. Zhang L, Wang H, Li C, Zhao Y, Wu L, Du X, et al. VEGF-A/Neuropilin 1 pathway confers cancer stemness via activating Wnt/Î²-catenin axis in breast cancer cells. *Cell Physiol Biochem*. (2017) 44:1251–62. doi: 10.1159/000485455
26. Palodetto B, Duarte ADSS, Lopes MR, Corrocher FA, Roversi FM, Niemann FS, et al. SEMA3A partially reverses VEGF effects through binding to neuropilin-1. *Stem Cell Res*. (2017) 22:70. doi: 10.1016/j.scr.2017.05.012
27. Zhigang Z, Bing P, Shaocheng L, Zhiwei J, Qian W, Ren L, et al. Integrating MicroRNA expression profiling studies to systematically evaluate the diagnostic value of MicroRNAs in pancreatic cancer and validate their prognostic significance with the cancer genome atlas data. *Cell Physiol Biochem*. (2018) 49:678–95. doi: 10.1159/000493033
28. Wu B, Wang K, Fei J, Bao Y, Wang X, Song Z, et al. Novel three - lncRNA signature predicts survival in patients with pancreatic cancer. *Oncol Rep*. (2018) 40:3427–37. doi: 10.3892/or.2018.6761
29. Xu Q, Gao J, Li Z. Identification of a novel alternative splicing transcript variant of the suppressor of fused: relationship with lymph node metastasis in pancreatic ductal adenocarcinoma. *Int J Oncol*. 49:2611–9. doi: 10.3892/ijo.2016.3753
30. Daoyan W, Liwei W, Masashi K, Zhiliang J, Xiangdong L, Qiang L, et al. KLF4α up-regulation promotes cell cycle progression and reduces survival time of patients with pancreatic cancer. *Gastroenterology*. (2010) 139:2135–45. doi: 10.1053/j.gastro.2010.08.022
31. Ueda J, Matsuda Y, Yamahatsu K, Uchida E, Naito Z, Korc M, et al. Epithelial splicing regulatory Protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. *Oncogene*. (2013) 33:4485–95. doi: 10.1038/onc.2013.392
32. Arafat H, Lazar M, Salem K, Chipitsyna G, Gong Q, Pan TC, et al. Tumor-specific expression and alternative splicing of the COL6A3 gene in pancreatic cancer. *J Surg Res*. (2011) 165:176. doi: 10.1016/j.jss.2010.11.782
33. Gottgens EL, Span PN, Zegers MMP. Roles and regulation of epithelial splicing regulatory proteins 1 and 2 in epithelial-mesenchymal transition. *Int Rev Cel Mol Bio*. (2016) 327:163–94. doi: 10.1016/bs.ircmb.2016.06.003
34. Tsai JH, Jing Y. Epithelial-mesenchymal plasticity in carcinoma metastasis. *Genes Dev*. (2013) 27:2192–206. doi: 10.1101/gad.225334.113
35. Warzecha CC, Shen S, Xing Y, Carstens RP. The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. *RNA Biol*. (2009) 6:546–62. doi: 10.4161/rna.6.5.9606
36. Francesca DM, Pierluigi I, Aaron B, Marcella M, Irene T, Letizia P, et al. Splicing program of human MENA produces a previously undescribed isoform associated with invasive, mesenchymal-like breast tumors. *Proc Natl Acad Sci USA*. (2012) 109:19280–5. doi: 10.1073/pnas.1214394109
37. Miwa T, Nagata T, Kojima H, Sekine S, Okumura T. Isoform switch of CD44 induces different chemotactic and tumorigenic ability in gallbladder cancer. *Int J Oncol*. (2017) 51:771–80. doi: 10.3892/ijo.2017.4063
38. Walser TC, Jing Z, Tran LM, Lin YQ, Yakobian N, Wang G, et al. Silencing the snail-dependent RNA splice regulator ESRP1 drives malignant transformation of human pulmonary epithelial cells. *Cancer Res*. (2018) 78:1986–99. doi: 10.1158/0008-5472.CAN-17-0315

39. Zong Z, Li H, Yi C, Ying H, Zhu Z, Wang H. Genome-wide profiling of prognostic alternative splicing signature in colorectal cancer. *Front Oncol*. (2018) 8:537. doi: 10.3389/fonc.2018.00537
40. Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KCH, et al. Multilayered control of alternative splicing regulatory networks by transcription factors. *Mol Cell*. (2017) 65:539–53. doi: 10.1016/j.molcel.2017.01.011
41. Sue N, Jack BH, Eaton SA, Pearson RC, Funnell AP, Turner J, et al. Targeted disruption of the basic Kruppel-like factor gene (Klf3) reveals a role in adipogenesis. *Mol Cell Biol*. (2008) 28:3967–78. doi: 10.1128/MCB.01942-07
42. Dong J, Chen C. Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases. *Cell Mol Life Sci*. (2009) 66:2691. doi: 10.1007/s00018-009-0045-z
43. Fan L, Hsieh PN, Sweet DR, Jain MK. Krüppel-like factor 15: regulator of BCAA metabolism and circadian protein rhythmicity. *Pharmacol Res*. (2018) 130:S1043661817314986. doi: 10.1016/j.phrs.2017.12.018
44. Guan F, Kang Z, Zhang J, Xue N, Yin H, Wang L, et al. KLF7 promotes polyamine biosynthesis and glioma development through transcriptionally activating ASL. *Biochem Bioph Res Commun*. (2019) 514:51–7. doi: 10.1016/j.bbrc.2019.04.120
45. Ding X, Wang X, Gong Y, Ruan H, Sun Y, Yu Y. KLF7 overexpression in human oral squamous cell carcinoma promotes migration and epithelial-mesenchymal transition. *Oncol Lett*. (2017) 13:2281–9. doi: 10.3892/ol.2017.5734
46. Shin SH, Cheol Kim S, Hong S, Kim YH, Song KB, Park K, et al. Genetic alterations of K-ras, p53, c-erbB-2, and DPC4 in pancreatic ductal adenocarcinoma and their correlation with patient survival. *Pancreas*. (2013) 42:216–22. doi: 10.1097/MPA.0b013e31825b6ab0
47. Makohonmoore AP, Zhang M, Reiter JG, Bozic I, Allen B, Kundu D, et al. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat Genet*. (2017) 49:358–66. doi: 10.1038/ng.3764
48. Bittoni A, Piva F, Santoni M, Andrikou K, Conti A, Loretelli C, et al. KRAS mutation status is associated with specific pattern of genes expression in pancreatic adenocarcinoma. *Future Oncol*. (2015) 11:1905–17. doi: 10.2217/fon.15.98
49. Tian J, Liu Y, Zhu B, Tian Y, Zhong R, Chen W, et al. SF3A1 and pancreatic cancer: new evidence for the association of the spliceosome and cancer. *Oncotarget*. (2015) 6:37750–7. doi: 10.18632/oncotarget.5647

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 September 2019
doi: 10.3389/fgene.2019.00879

Comprehensive Analysis of Expression and Prognostic Value of Sirtuins in Ovarian Cancer

Xiaodan Sun[1,2], Shouhan Wang[3] and Qingchang Li[1,4]**

[1] Department of Pathology, College of Basic Medical Sciences, China Medical University, Shenyang, China, [2] Department of 2nd Gynecologic Oncology Surgery, Jilin Cancer Hospital, Changchun, China, [3] Department of Hepatopancreatobiliary Surgery, Jilin Cancer Hospital, Changchun, China, [4] Department of Pathology, the First Affiliated Hospital, China Medical University, Shenyang, China

Sirtuins (SIRTs) 1–7 are a family of intracellular enzymes, which possess nicotinamide adenine dinucleotide-dependent deacetylase activity. Emerging evidence suggest that SIRTs play vital roles in tumorigenesis by regulating energy metabolism, DNA damage repair, genome stability, and other cancer-associated cellular processes. However, the distinct roles of the seven members in ovarian cancer (OC) remain elusive. The transcriptional expression patterns, prognostic values, and genetic alterations of seven SIRTs in OC patients were investigated in this study using a range of databases: Oncomine and Gene Expression Profiling Interactive Analysis, Kaplan–Meier plotter, the Cancer Genome Atlas, and cBioPortal. The protein–protein interaction networks of SIRTs were assessed in the String database. Gene Ontology enrichment and Kyoto Encyclopedia of Genes and Genomes pathway were analyzed in Database for Annotation, Visualization, and Integrated Discovery. The mRNA expression levels of SIRT1–4 and 7 were downregulated, while that of SIRT5 was upregulated and SIRT6 exhibited both expression dysregulation in patients with OC. Dysregulated SIRTs mRNA expression levels were associated with prognosis. Moreover, genetic alterations primarily occurred in SIRT2, 5, and 7. Network analysis indicated that SIRTs and their 20 interactors were associated with tumor-related pathways. This comprehensive bioinformatics analysis revealed that SIRT1–4, 6, and 7 may be new prognostic biomarkers, while SIRT5 is a potential target for accurate therapy for patients with OC, but further studies are needed to confirm this notion. These findings will contribute to a better understanding of the distinct roles of SIRTs in OC.

Keywords: sirtuins, ovarian cancer, prognosis, database, bioinformatics analysis

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Shuangyu Lv,
Henan University, China
Nan Wu,
Peking Union Medical College Hospital (CAMS), China

****Correspondence:***
Shouhan Wang
15640584861@163.com
Qingchang Li
qcli@cmu.edu.cn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *12 June 2019*
Accepted: *21 August 2019*
Published: *13 September 2019*

Citation:
Sun X, Wang S and Li Q (2019) Comprehensive Analysis of Expression and Prognostic Value of Sirtuins in Ovarian Cancer. Front. Genet. 10:879. doi: 10.3389/fgene.2019.00879

INTRODUCTION

Ovarian cancer (OC) ranked eighth in incidence and seventh in mortality rates globally among all cancers in women in 2018 (WHO, http://gco.iarc.fr/today/home). Furthermore, the absence of incipient symptoms leads to over three quarters of patients being diagnosed at advanced stages (Zhou et al., 2018). Standard treatment for this disease involves surgical intervention combined with chemotherapy. Although the use of gene sequencing and targeted therapies have improved the survival of OC patients, the 5-year survival rate is still poor because of the complex tumor processes and pathological subtypes of OC and the shortage of more specific target biomarkers. Therefore,

enhancing therapy requires new biomarkers for prognosis and individualized treatment of OC.

Sirtuins (SIRTs) are a family of intracellular enzymes that possess nicotinamide adenine dinucleotide (NAD^+)-dependent deacetylase activity and share a highly conserved 275-amino catalytic core domain. Seven members (SIRT1–7) in mammals are divided into the following four classes: SIRT1–3, I; SIRT4, II; SIRT5, III; and SIRT6-7, IV (O'Callaghan and Vassilopoulos, 2017). Based on their subcellular localization, they can also be categorized as follows: SIRT1, 6, and 7 reside in the nucleus; SIRT2 is expressed in both the nucleus and cytoplasm; and SIRT3, 4, and 5 are in the mitochondria (Chalkiadaki and Guarente, 2015). Emerging evidence suggest that SIRTs play vital roles in tumorigenesis by regulating energy metabolism, DNA damage repair, genome stability, and various other cancer-associated cellular processes. Aberrant expression of SIRTs has been found in common human carcinomas such as breast, lung, liver, and gastrointestinal cancers, as well as OC and neurologic tumors (Chen et al., 2013; Chalkiadaki and Guarente, 2015; Osborne et al., 2016; O'Callaghan and Vassilopoulos, 2017).

Presently, the dysregulated expression of SIRTs and their prognostic value have been partly reported in OC. For example, the expression of SIRT1 was found to be higher in 68 OC tissue samples than it was in 16 normal ovaries (Mvunta et al., 2017). Consistent with this study, overexpression of SIRT1 was also reported in 90 OC tissue samples compared with 40 normal ovary tissues, and, interestingly, a high expression level of SIRT1 was associated with a favorable outcome (Jang et al., 2009). However, a converse finding that SIRT1 was downregulated in OC based on public datasets has also been reported (Hyde et al., 2018). SIRT2 predicted poor survival when upregulated in patients with OC (Teng and Zheng, 2017), while reduced expression of SIRT2 was observed in 13 samples of serous ovarian carcinoma compared with 11 samples of normal ovarian surface epithelial tissues (Du et al., 2017). At least one copy of the *SIRT3* gene was deleted in 40% of breast and OCs, and focal deletions of *SIRT3* were especially frequent in ovarian tumors (Finley et al., 2011). In contrast, the region encompassing the *SIRT5* locus was amplified in 30% of high-grade serous ovarian carcinomas (Bell et al., 2011a). SIRT3 and SIRT5 expression were found to be significantly decreased and increased in primary serous OCs/tubal cancers compared with that in normal counterparts, respectively (Li et al., 2019). SIRT4 has been reported to function as a tumor suppressor in published studies, and reduced expression in OC was reported in a meta-analysis (Csibi et al., 2013).The mRNA expression of SIRT6 in 32 OC tissue samples was remarkably lower than that in paired normal ovarian tissues (Zhang et al., 2015), whereas there were higher SIRT7 mRNA levels in OC, although without statistical significant, which could have been due to the small sample sizes analyzed (Aljada et al., 2015).

These findings indicate that SIRTs are closely associated with OC, and it is striking that even in the same tumor, the specific roles of individual SIRTs can be controversial, which may be partly ascribed to small sample sizes. A comprehensive analysis of the expression and mutation patterns and prognostic values of SIRTs in OC based on large database analysis would enhance the understanding of their potential roles in OC. Therefore, we conducted this study to investigate this phenomenon.

METHODS

Ethics Statement

The OC specimens and normal tissues were obtained from patients who were diagnosed with OC and underwent primary cytoreductive (debulking) surgery from Aug 2017 and May 2018 in First Affiliated Hospital, China Medical University. The enrolled patients had signed informed consent. This study was approved by the Medical Research Ethics Committee of China Medical University and conducted according to the principles expressed in the Declaration of Helsinki. All the datasets were retrieved from the published literature, so it was confirmed that all written informed consent was obtained.

Oncomine Database

The Oncomine database (www.oncomine.org) (Rhodes et al., 2004), an online cancer microarray database and web-based data-mining platform, was used to investigate the transcriptional levels of SIRTs in different clinical cancer specimens and corresponding normal controls. The search contents and thresholds were set as follows: keywords, SIRT1–SIRT7, primary filter, cancer vs. normal; cancer type, OC, the absolute value of $\log_2$ fold change >1.5, $P < 0.05$; and gene rank, 10%. The P value was calculated using the Student's t test.

GEPIA Database

The Gene Expression Profiling Interactive Analysis (GEPIA) database (http://gepia.cancer-pku.cn/), a newly developed web-based tool, provides key interactive and customizable functions including tumor vs. normal differential expression analysis, profiling plotting in accordance with cancer types or different pathological stages, correlation analysis, patient survival analysis, similar gene detection, and dimensionality reduction analysis based on the Cancer Genome Atlas (TCGA) and the genotype–tissue expression data (Tang et al., 2017).

The Kaplan–Meier Plotter

The prognostic value of SIRTs in OC patients was evaluated using the Kaplan–Meier plotter (http://kmplot.com/analysis), an open online dataset that can be used to assess the effect of 54,675 genes on survival in 21 cancer types including breast, liver, ovarian, lung, and gastric cancer (Győrffy et al., 2012). To analyze the overall survival (OS) and progression-free survival (PFS) of patients with OC, samples were split into two groups based on median expression (high vs. low). The hazard ratio (HR) with 95% confidence intervals (CIs) and log-rank P values were calculated and displayed in survival plots. $P < 0.05$ was considered statistically significant.

TCGA Database and cBioPortal

The cBioPortal for Cancer Genomics (http://cbioportal.org) provides an open-access web resource for exploring, visualizing,

and analyzing multidimensional cancer genomic data from TCGA (Gao et al., 2013). In the present study, three TCGA datasets of OC, namely, "TCGA Nature 2011 (563 cases)," "TCGA PanCancer Atlas (585 cases)," and "TCGA Provisional (606 cases)" were selected for further analysis of *SIRT* gene mutations or copy number alterations (CNA). The OncoPrint, survival tabs were applied according to the online instructions of the cBioPortal.

String Database and DAVID

The interaction proteins network of SIRTs was constructed using the String Database (https://string-db.org/), which is an online database of predicted functional associations between proteins (von Mering et al., 2003). "*Homo sapiens*" was selected and interactions with a combined score >0.7 (high confidence) were considered significant. Seven SIRTs and 20 associate proteins were imported into Database for Annotation, Visualization, and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/) to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses (Huang et al., 2009a; Huang et al., 2009b). The human genome was selected as the background parameter, and a $P < 0.05$ was considered statistically significant.

Immunohistochemistry

Surgically excised normal and tumor specimens were fixed in 10% neutral formalin, embedded in paraffin, and cut into 4-mm sections. The sections were incubated with commercial rabbit polyclonal antibodies against SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, and SIRT7 (SIRT1, 2, 5–7 were purchased from Proteintech, China; SIRT3 and SIRT4 were purchased from Abcam, China) at 1/100 dilution overnight at 4°C. Then, the reaction was visualized using the Elivision super HRP IHC Kit (Maixin-Bio) and 3,3-diaminobenzidine (DAB); nuclei were counterstained with hematoxylin. The sections were dehydrated in ethanol before mounting.

Cell Culture and Quantitative Real-Time PCR Analysis

The A2780 and SKOV-3 human OC cell lines were used in this study. The cells were cultured in Dulbecco's modified Eagle medium and RPMI-1640, respectively, supplemented with 10% fetal bovine serum. These cells were grown at 37°C in a humidified atmosphere with 5% CO_2.

Trizol (Invitrogen, Carlsbad, CA) was used to extract total RNA from OC cells. One microgram RNA was reverse transcripted using the PrimeScript RT Master Mix (TaKaRa) according to manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) was done using Applied Biosystems Power SYBR Green on a qTOWER2.0. Real-time PCR system is as follows: 10 s at 95°C, then 40 cycles at 95°C for 5 s, and 65°C for 34 s. The gene amplification specificity was shown by a melting curve generated in dissociation procedure. $2^{-\Delta\Delta Ct}$ method was used to normalize the quantification of SIRT1-7 to glyceraldehyde 3-phosphate dehydrogenase (GAPDH). The specific primer sequences are performed as follows:

GAPDH Forward	Reverse
5′-CCACCCATGGCAAATTCC-3′	5′-GATGGGATTTCCATTGATGACA-3′
SIRT1 Forward	Reverse
5′-GTAGGCGGCTTGATGGTAATC-3′	5′-GACTCTGGCATGTCCCACTAT-3′
SIRT2 Forward	Reverse
5′-GCGGAACTTATTCTCCCAGAC-3′	5′-GCTCCCACCAAACAGATGAC-3′
SIRT3 Forward	Reverse
5′-CTGTGGGTGCTTCAAGTGTTG-3′	5′-CCCGAATCAGCTCAGCTACAT-3′
SIRT4 Forward	Reverse
5′- ACACTGGGCTTTGAGCACCT-3′	5′-GAGTCTGTTCCCCACAATCCA -3′
SIRT5 Forward	Reverse
5′-TCGTGGTCATCACCCAGAAC-3′	5′-GCCACAACTCCACAAGAGGTAC-3′
SIRT6 Forward	Reverse
5′-GCCAAGTGTAAGACGCAGTAC-3′	5′-TAGGATGGTGTCCCTCAGCT-3′
SIRT7 Forward	Reverse
5′-CATCGTGAACCTGCAGTGGA-3′	5′-GGGAGTCGCCAGTGAGAAAA-3′

RESULTS

Transcriptional Levels of SIRTs and Their Relationship With Clinicopathological Characters in Patients With OC

The dysregulated transcriptional levels of seven SIRTs have been identified in 20 different types of human cancers in the Oncomine database. As shown in **Figure 1**, SIRTs might act as either a tumor promoter or suppressor, in a context-specific manner. Especially, the mRNA expression levels of SIRT1 were significantly downregulated in patients with OC in Bonome's dataset (Bonome et al., 2008) with a $\log_2$ fold change of −1.866, while SIRT5 and SIRT7 were higher in ovarian serous adenocarcinoma in two another datasets (Yoshihara Ovarian and TCGA datasets; $\log_2$ fold changes, 1.929 and 1.626, respectively) (Yoshihara et al., 2009) than in normal ovarian tissues (**Table 1**, bold font).

Moreover, the mRNA levels of SIRTs in different types of OC, which were available in Oncomine datasets, are summarized in **Table 1**. In Hendrix's dataset, SIRT1, SIRT3, and SIRT4 expression levels were significantly lower in serous, endometrioid, mucinous, and clear cell adenocarcinoma than they were in normal ovarian tissues. SIRT2 expression was lower in serous and endometrioid adenocarcinoma in Lu's dataset (Lu et al., 2004), whereas SIRT5 was upregulated in those types of OC in Hendrix's dataset compared with normal tissues (Hendrix et al., 2006). SIRT6 was expressed at higher levels in all types of OC than it was in normal tissues in Hendrix's dataset except for serous adenocarcinoma. Interestingly, SIRT7 was downregulated in OC in Bonome's dataset but upregulated in both TCGA and Hendrix's datasets compared with normal tissues (Hendrix et al., 2006; Bonome et al., 2008).

In addition, the GEPIA database was also used to compare the mRNA expression of SIRTs between OC and normal tissues. The expression levels of SIRT1–3 were significantly lower, and levels of SIRT4, 6, and 7 were slightly more downregulated ($P > 0.05$) in OC than they were in normal tissues, while SIRT5 exhibited contrasting expression (**Figure 2A**). The results were

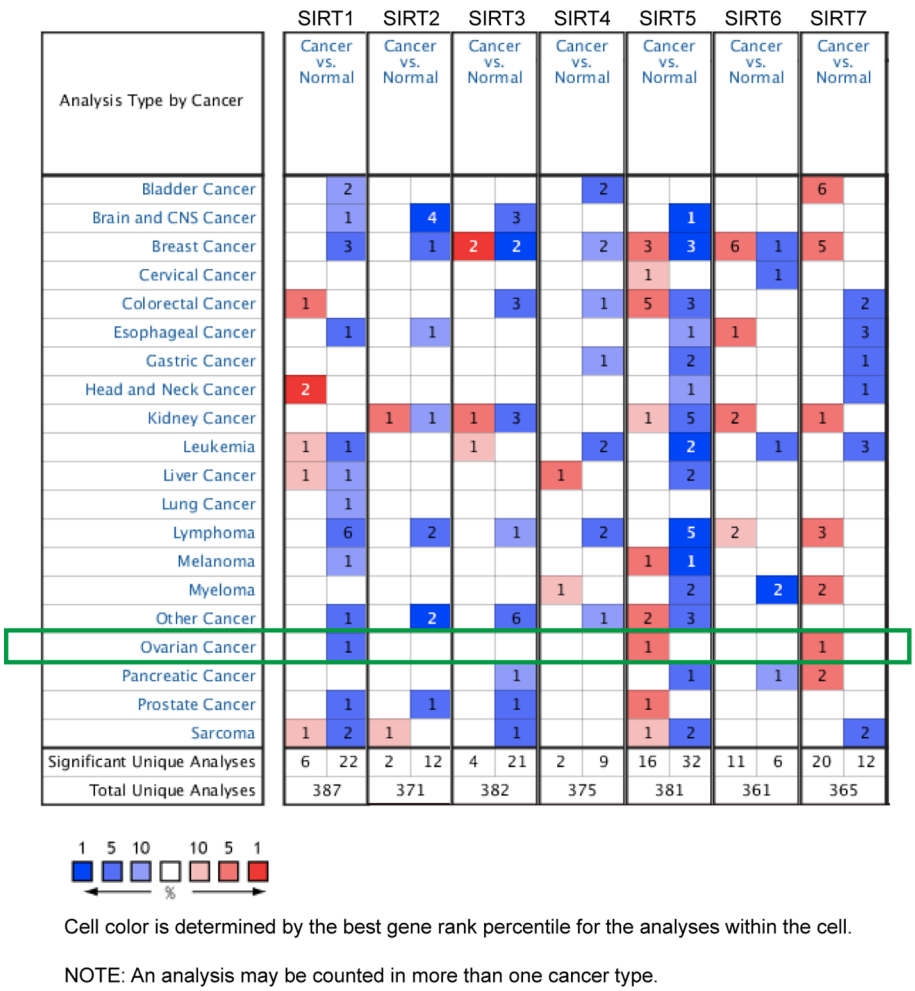


FIGURE 1 | The mRNA levels of sirtuins (SIRTs) in 20 different types of cancers (Oncomine). The number in each cell represents the number of analyses that satisfied the following threshold: $P < 0.05$, the absolute value of $\log_2$ fold change >1.5, and gene rank, 10%. The numbers in colored cells show the quantities of datasets with statistically significant mRNA overexpression (red) or downexpression (blue) of target genes.

consistent with those of the Oncomine database except for that of SIRT6. These findings were verified by immunochemistry (IHC), and as shown in **Figure 2B**, SIRT5 protein expression was higher in OC than in the counterpart normal tissues, while the protein expression difference of other SIRTs was not significant. Furthermore, the mRNA levels of SIRTs in two OC cell lines were detected by qRT-PCR, and the results were similar to the IHC (**Figure 2C**). The relationship between mRNA expression levels of SIRTs and different tumor stages of OC were also analyzed, and they were all significantly upregulated in stage II except for SIRT2 and SIRT4 (**Figure 3**).

Prognostic Value of SIRTs in Patients With OC

To further assess the prognostic value of SIRTs in all patients with OC, Kaplan–Meier plotter analysis was used. We initially assessed the relationship between the mRNA expression of individual SIRT and the survival of OC patients. The survival curves demonstrated that decreased SIRT1 and SIRT4 mRNA levels and increased expression of SIRT2, 3, 6, and 7 predicted favorable prognosis (OS and PFS). Interestingly, a higher level of SIRT5 was associated with shorter PFS but with longer OS. Then, we also wondered the prognostic value of the combined SIRTs, and the results showed that upregulated levels of their combined mRNA expression was correlated with poor outcome in patients with OC (**Figure 4**).

Moreover, we also assessed the prognostic values of SIRTs in different subtypes of OC, namely, different histology, clinical stages, pathological grades, and TP53 status, which are available in Kaplan–Meier plotter. As shown in **Table 2**, increased mRNA expression of SIRT3, 5, 6, and 7 in serous OC patients and decreased levels of SIRT4 in both serous and endometrioid OC patients were significantly related to improved OS. The overexpression of SIRT2–4 predicted shorter PFS in serous OC patients. As shown in **Table 3**, high mRNA expression of SIRT5 and low expression of SIRT6, 7 were associated with poor OS in stage 1. Elevated mRNA levels of SIRT3, 5–7 and low levels of

TABLE 1 | The significant changes of sirtuin (SIRT) expression between different types of OC and normal tissues (Oncomine).

Sirtuins	Types of OC vs. Normal												Ref/Source
	Ovarian Carcinoma			Serous			Endometrioid			Mucinous			Clear cell			
	Log_2 FC	*P*	N	Log_2 FC	*P*	N	Log_2 FC	*P*	N	Log_2 FC	*P*	N	Log_2 FC	*P*	N	
SIRT1	**−1.866**	**1.19E-9**	185	–	–	–	–	–	–	–	–	–	–	–	–	Bonome ovarian
	–	–	–	−1.172	1.67E-5	41	−1.142	4.92E−5	37	−1.174	1.62E−5	13	−1.128	0.001	8	Hendrix ovarian
SIRT2	–	–	–	−1.146	0.021	20	−1.179	0.011	9	NS	NS	9	NS	NS	7	Lu ovarian
SIRT3	–	–	–	−1.151	0.004	41	−1.150	0.005	37	−1.159	0.013	13	−1.153	0.012	8	Hendrix ovarian
SIRT4	–	–	–	−1.177	3.30E-6	41	−1.171	4.82E−6	37	−1.179	9.29E−4	13	−1.074	0.049	8	Hendrix ovarian
SIRT5	–	–	–	1.115	1.29E-4	41	1.041	0.033	37	NS	NS	13	NS	NS	8	Hendrix ovarian
	–	–	–	**1.929**	**6.44E-7**	43	–	–	–	–	–	–	–	–	–	Yoshihara ovarian
SIRT6	–	–	–	NS	NS	41	1.033	0.007	37	1.056	5.15E−4	13	1.070	0.002	8	Hendrix ovarian
SIRT7	−1.097	0.025	185	–	–	–	–	–	–	–	–	–	–	–	–	Bonome ovarian
	–	–	–	**1.626**	**1.71E-8**	586	–	–	–	–	–	–	–	–	–	TCGA
	–	–	–	1.163	1.95E-8	41	1.15	8.14E−8	37	1.255	1.45E−4	13	1.233	1.33E−4	8	Hendrix ovarian

The bold font indicates the difference between OC and normal tissues meets the selected thresholds.

OC, ovarian cancer; FC, fold change; NS, not significant; "–", not available; N, number of patients.

SIRT1, 4 were associated with better OS in stage 3, while high level of SIRT2 predicted poor OS in stage 4. In terms of pathological grades, high SIRT6 mRNA expression was linked to favorable OS. Interestingly, increased expression of SIRT3 predicted poor OS in mutated TP53 type, while it was associated with better OS in wild-type TP53. With respect to PFS (**Table 4**), high mRNA expression of SIRT1-3 and 7 were found to be correlated to shorter PFS in stage 1, whereas low levels of SIRT1, 5 and SIRT2, 4, and 6 predicted longer PFS in stages 2 and 3, respectively. In stage 4, increased expression of SIRT2 and 3 were linked to poor PFS. With regard to pathological grades, decreased levels of SIRT2 and 4 predicted better PFS. Interestingly, SIRT3 exhibited opposite roles in different pathological grades. Additionally, elevated expression of SIRT1 and 2 were associated with poor PFS in both mutated and wild type of TP53, while increased levels of SIRT3, 6, and 7 were related to poor PFS in mutated TP53 status. Taken together, these results indicated that the mRNA expression levels of SIRTs may be potential biomarkers for the prediction of OC patient survival.

Genetic Alteration Analysis of SIRTs in Patients With OC

Next, the genetic alterations of SIRTs in OC patients were explored using the TCGA database and c-BioPortal online tool. SIRTs were altered in 1,754 samples of 1,742 patients from three TCGA databases of serous cystadenocarcinoma, and the alteration rates were 31.02% (188/606), 24.1% (141/585), and 16.7% (94/563), respectively, and the amplification accounted for most changes (**Figure 5A**). As shown in **Figure 5B**, the genetic SIRT alterations occurred in 423 (24%) of the queried samples, and the individual sequence alteration rates varied from 1.4 to 10%. SIRT2, SIRT5, and SIRT7 were ranked as the top 3 of the seven members, and their mutation rates were 10, 8, and 5%, respectively (**Figure 5B**). Using the "Survival" tab with the Kaplan–Meier plot and log-rank test, the survival curves showed that cases with or without alterations in one of the SIRTs had no relationship with OS and PFS (**Figures 5C**, **D**).

GO Enrichment and KEGG Pathway Analysis of Protein–Protein Interaction of SIRTs

A network of seven SIRT members and 20 proteins that significantly interacted with SIRTs was constructed using the String database [protein–protein interaction (PPI) enrichment $P < 1.0E{-}16$]. The network graphic showed that cell metabolism-related genes tumor protein 53 (*TP53*), Fork head box O 1/3/4 (*FOXO1/3/4*), and superoxide dismutase 2 (*SOD2*), and histone posttranscriptional modification-related genes histone deacetylase 1/2/4 (*HDAC 1/2/4*), E1A binding protein p300 (*EP300*), and suppressor of variegation 3–9 homolog 1 (*SUV39H1*) were associated with SIRTs (**Figure 6A**). Then, using "correlation analysis" in GEPIA, the Pearson correlation coefficients were calculated between SIRTs (**Figure 6B**), ranging from 0.073 (SIRT1 vs. SIRT2) to 0.39 (SIRT1 vs. SIRT3).

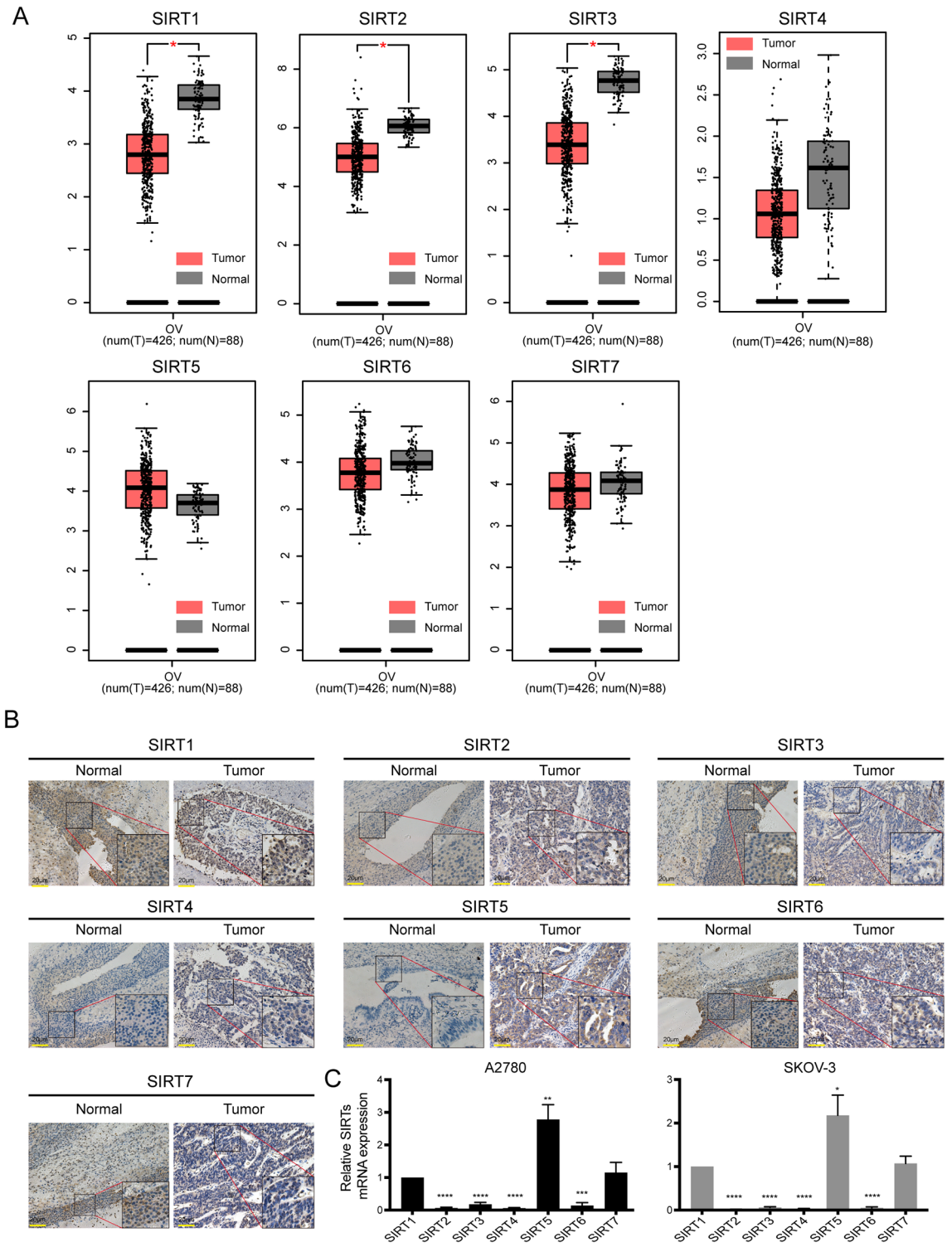


FIGURE 2 | The mRNA and protein expression of SIRTs in patients with ovarian cancer (OC). **(A)** Box plots of SIRTs mRNA expression based on GEPIA database. **(B)** The representative immunohistochemical staining images of SIRTs protein expression in ovarian cancer and normal tissues (magnification, ×400; scale bar = 20 μm). **(C)** The mRNA levels of SIRTs in A2780 and SKOV-3 ovarian cell lines by quantitative real-time PCR (qRT-PCR).*$P < 0.05$, **$P < 0.01$,***$P < 0.001$, ****$P < 0.00001$.

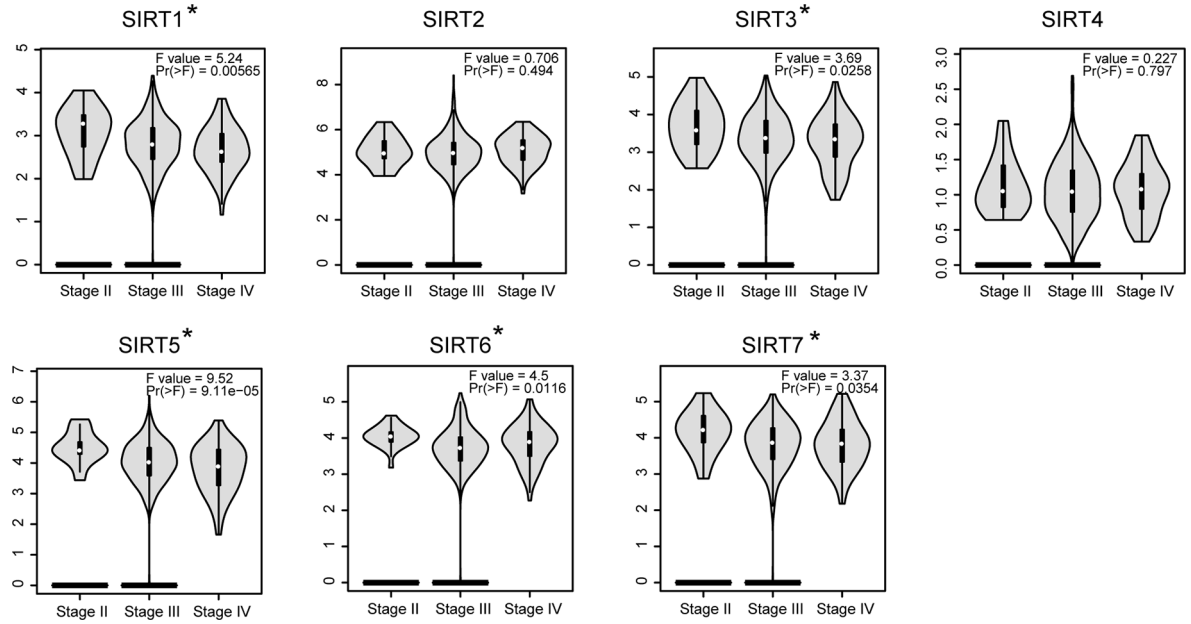


FIGURE 3 | The relationship between SIRTs mRNA expression and tumor stages in patients with OC [Gene Expression Profiling Interactive Analysis (GEPIA)]. *$P < 0.05$.

Next, GO enrichment and KEGG pathway analysis of SIRTs and their interactors were performed using DAVID. Cellular components, biological process, and molecular functions were the three main functions of target host genes in the GO enrichment analysis. The nucleoplasm, nucleus, and cytoplasm were the major cellular components of target genes (**Figure 7A**). Regulation of transcription from RNA polymerase II promoter and DNA templated were mainly associated with SIRTs and their interacting neighbors while binding to DNA, chromatin, and transcription factor were their primary molecular functions predicted online (**Figures 7B**, **C**). The top 10 KEGG pathways for target genes are shown in **Figure 7D**, and the Notch, FOXO, and cancer pathways were found to be invoved in OC.

DISCUSSION

Emerging evidence suggest that SIRTs play vital roles in tumorigenesis mediated by their ability to regulate energy metabolism, DNA damage repair, genome stability, and other cancer-associated cellular processes. However, the distinct roles of seven SIRT members in OC are yet to be elucidated. In the current study, the mRNA expression patterns, prognostic values, genetic alterations, and PPI networks of SIRTs in OC patients were investigated through various large databases, including Oncomine and GEPIA, Kaplan–Meier Plotter, cBioPortal, and String. Moreover, GO enrichment and KEGG pathway were also analyzed *via* DAVID.

SIRT1 is the most studied of these seven SIRT members in human cancer and plays dual roles in numerous malignancies including OC (Chalkiadaki and Guarente, 2015). For example, the expression of SIRT1 was significantly higher in endometrioid, mucinous, and clear-cell OC than it was in normal ovaries in IHC analysis, and its overexpression predicted shorter survival in OC (Mvunta et al., 2017). Moreover, overexpression of nuclear SIRT1 was also found to induce chemoresistance and poor prognosis in 63 OC patients (Shuang et al., 2015). Consistently, SIRT1 was found to be involved in the high expression of cancer stem cell markers, chemoresistance, tumorigenesis, and epithelial to mesenchymal transition (EMT) phenotype (Qin et al., 2017). In contrast to these findings, SIRT1 was downregulated in OC based on public datasets and acts as a tumor suppressor (Hyde et al., 2018). In our study, the mRNA expression of SIRT1 was markedly lower in OC tissues than it was in normal tissues. Interestingly, a higher mRNA expression of SIRT1 was significantly associated with poor outcome in OC.

SIRT2 was initially implicated in mitotic progression and serves as a cell cycle regulator (Dryden et al., 2003). Recently, several studies have highlighted the critical roles of SIRT2 in maintaining genome stability (Kim et al., 2011; Serrano et al., 2013), suggesting that this SIRT mainly functions as a tumor suppressor (Chalkiadaki and Guarente, 2015). For example, SIRT2 expression in serous OC was significantly lower than it was in ovarian surface epithelium as determined using Western blotting and IHC. Reduced expression of SIRT2 upregulated cyclin-dependent kinase 4 (CDK4) expression, which eventually accelerated cell proliferation, migration, and invasion, indicating that SIRT2 plays a tumor-suppressor role in OC (Du et al., 2017). Consistently, in the present study, the mRNA expression of SIRT2 was considerably more decreased in OC, especially serous and endometrioid subtypes, than it was in normal tissues and increased levels predicted favorable OS and PFS in patients with OC. However, overexpression of SIRT2 was previously reported to have been related to a poor prognosis in 491 patients with OC

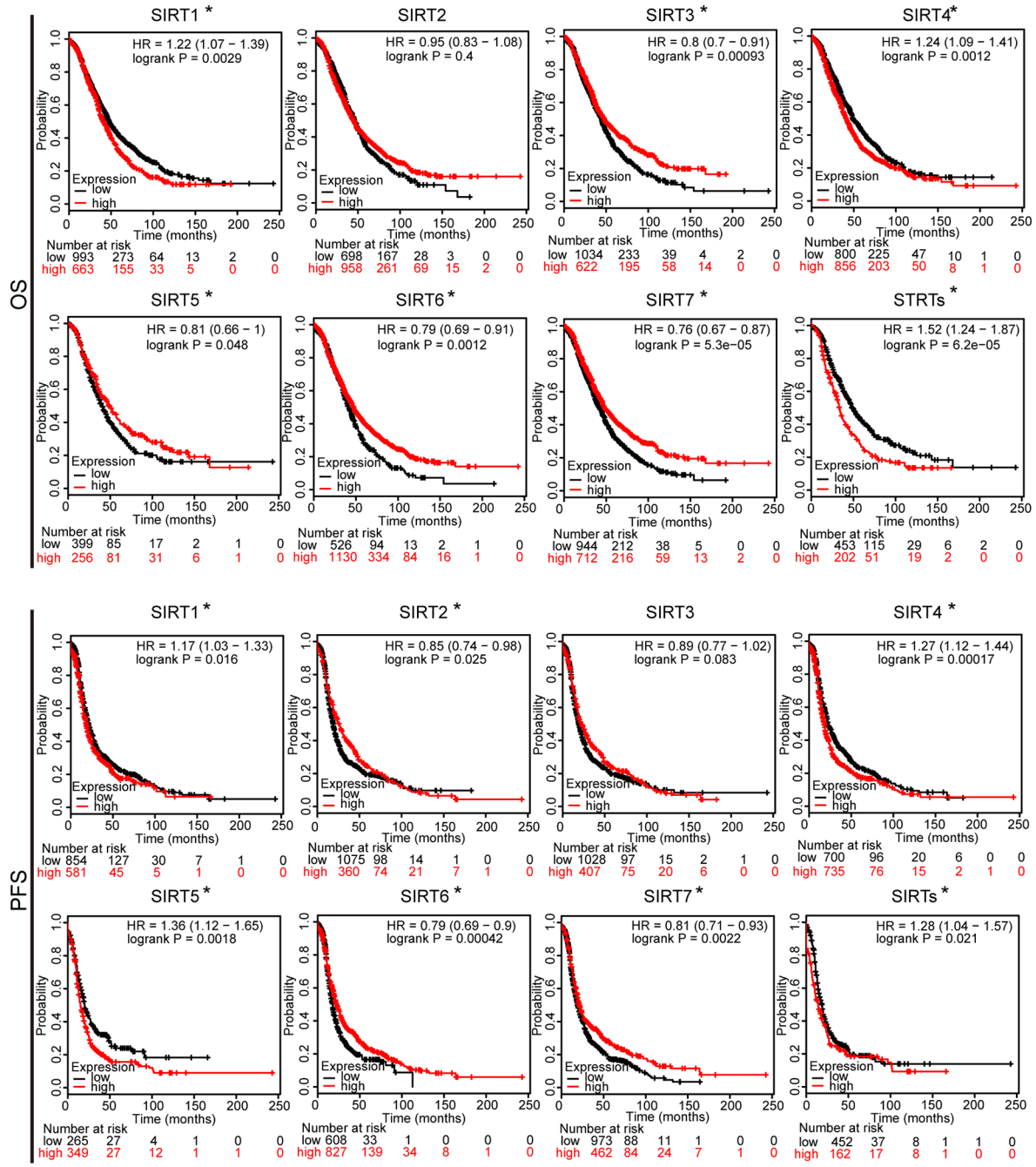


FIGURE 4 | The prognostic value of mRNA level of SIRTs in patients with OC (Kaplan–Meier plotter). *$P < 0.05$.

(Teng and Zheng, 2017). We assumed that this discrepancy may be due to the high mutation rate of *SIRT2* (10%) in OC, which was identified in our study.

SIRT3 primarily serves as a tumor suppressor by limiting reactive oxygen species levels and antagonizing hypoxia-inducible factor 1-α, which fights against a metabolic switch to aerobic glycolysis (Bell et al., 2011b; Finley et al., 2011; Chalkiadaki and Guarente, 2015). SIRT3 was reported to be downregulated in both metastatic tissues and cell lines of OC and inhibit EMT by interacting with and repressing Twist (Xiang et al., 2016). Moreover, SIRT3 was reported to be activated by S1, a novel pan B-cell lymphoma-2 inhibitor, and then it exerted a proapoptotic effect in SKOV3 OC cells (Dong et al., 2016). SIRT3 was identified to decrease and function as an independent favorable prognostic factor for OS in serous OC (Li et al., 2019). Similarly, our study demonstrated that the transcription levels of SIRT3 in different subtypes of OC were remarkably lower than those in normal samples, and its increased mRNA expression was significantly

TABLE 2 | The prognostic values of SIRTs in different pathological subtypes OC (Kaplan–Meier plotter).

Sirtuins	Histology	OS			PFS		
		Cases	HR(95% CI)	*P* value	Cases	HR(95% CI)	*P* value
SIRT1	Serous	1,207	1.15(0.99–1.34)	0.074	1,104	0.88(0.75–1.03)	0.1
218878_s_at	Endometrioid	37	4.94(0.82–29.69)	0.053	51	0.56(0.22–1.43)	0.22
SIRT2	Serous	1,207	1.13(0.95–1.33)	0.17	1,104	1.4(1.2–1.63)	**1.6E–05**
220605_s_at	Endometrioid	37	3.84(0.43–34.41)	0.10	51	2.08(0.82–5.27)	0.11
SIRT3	Serous	1,207	0.82(0.7–0.95)	**0.0096**	1104	1.21(1.03–1.41)	**0.019**
221913_at	Endometrioid	37	0.46(0.08–2.75)	0.38	51	4.92(0.65–36.99)	0.086
SIRT4	Serous	1,207	1.22(1.05–1.42)	**0.011**	1104	1.26(1.09–1.45)	**0.0019**
220047_at	Endometrioid	37	9.36(1.04–84.6)	**0.016**	51	0.64(0.21–1.94)	0.42
SIRT5	Serous	523	0.78(0.62–0.98)	**0.036**	483	1.17(0.94–1.47)	0.17
229112_at	Endometrioid	30	3.01(0.31–29)	0.32	44	1.51(0.47–4.83)	0.48
SIRT6	Serous	1,207	0.81(0.69–0.94)	**0.0062**	1104	1.14(0.97–1.33)	0.11
219613_s_at	Endometrioid	37	0.17(0.02–1.5)	0.069	51	1.97(0.7–5.55)	0.19
SIRT7	Serous	1,207	0.8(0.69–0.93)	**0.0044**	1104	1.1(0.93–1.3)	0.28
218797_s_at	Endometrioid	37	–	0.18	51	1.99(0.79–5.03)	0.14

The bold font indicates the difference was significant statistically. "–", not available; OC, ovarian cancer; OS, overall survival; PFS, progression-free survival.

associated with tumor stage II and favorable outcome in OC. In addition, our results showing that the genetic alteration rate of *SIRT3* was 2.4% and extensive deletion predominately occurred were in line with the findings that at least one copy of the *SIRT3* gene was deleted in 40% of breast cancers and OC, and focal deletions of *SIRT3* were especially frequent (Finley et al., 2011).

SIRT4 has been largely reported to have protective roles against cancer by repressing glutamine metabolism and maintaining genomic stability (Fernandez-marcos and Serrano, 2013; Chalkiadaki and Guarente, 2015). However, its expression pattern and prognostic value in OC have been rarely reported. Only one meta-analysis suggested that lower expression of the *SIRT4* gene was found in a series of solid carcinomas including OC than in corresponding normal tissue (Csibi et al., 2013). Likewise, our results showed that a lower mRNA expression of SIRT4 was found in OC than in normal tissues. Interestingly, a decreased level of SIRT4 was associated with unfavorable OS and PFS in OC, especially in serous subtypes. Although it is not clear, we ascribed the contradictory findings to the background heterogeneity between different databases.

SIRT5 is a unique member of the SIRT family, which possesses multiple enzymatic activities including NAD-dependent histone deacetylase (Nakagawa et al., 2009), potent lysine demalonylase, desuccinylase (Du et al., 2011), and lysine glutarylase (Tan et al., 2014), now known to play controversial roles in tumorigenesis. However, an understanding of the distinct role of SIRT5 in OC is still in its infancy. An analysis of human high-grade serous ovarian carcinomas revealed that the region encompassing the *SIRT5* locus was amplified in 30% of these tumors (Bell et al., 2011a). Consistently, our results showed *SIRT5* gene alteration in 8% of queried OC patients and amplifications accounted for most CNAs. Moreover, SIRT5 was found to increase in primary serous OCs/tubal cancers compared with that in normal tissues, and high expression of it was associated with better OS by univariable analysis (Li et al., 2019). Similarly, in our study, a higher mRNA level of SIRT5 was found in OC, especially in serous adenocarcinoma, and it was related to poor PFS in OC. Interestingly, increased expression of SIRT5 predicted superior OS, and this may be partly due to its marked overexpression in early tumor stages.

SIRT6 and SIRT7 are both nuclear proteins with deacetylase activity and function as both tumor suppressor and promotor in cancer, including OC (Chen et al., 2013; Chalkiadaki and Guarente, 2015). The mRNA expression of SIRT6 in 32 OC tissue samples was remarkably lower than that in the paired normal tissues, and SIRT6 inhibited the proliferation of OC cells by suppressing Notch 3 expression (Zhang et al., 2015). Conversely, the expression of SIRT6 was associated with higher tumor stage, higher histological grade, platinum resistance, and predicted shorter OS in 104 patients with OC. Moreover, SIRT6 was overexpressed in omental metastases compared with corresponding primary counterparts (Li et al., 2019) and facilitated the invasiveness of OC cells by regulating EMT signaling, but it did not inhibit their proliferation (Bae et al., 2018).

SIRT7 was overexpressed in OC tissues and cell lines (Barber et al., 2013), omental metastasis tissues (Li et al., 2019), and promoted tumor cell proliferative potential *via* regulating apoptosis (Wang et al., 2015). However, SIRT7 was significantly reduced in cultured chemoresistant OC cells (Aljada et al., 2014) and was considered a tumor suppressor based on its inhibition of the activity of HIF-1 and HIF-2 transcription factors (Hubbi et al., 2013). The present study demonstrated that SIRT6 and SIRT7 levels were slightly lower in OC than normal conditions based on the GEPIA database analysis ($P > 0.05$) but significantly upregulated in the Oncomine database. Moreover, overexpression of SIRT6 and SIRT7 was associated with tumor stage II and a better outcome.

In addition to the individual prognostic values of the investigated SIRTs, we further determined the simultaneous increase in the mRNA expression of all SIRTs predicted poor prognosis and whether the genes altered or not had no relationship with OS and PFS. In addition, the enrichment analysis indicated

TABLE 3 | The relationship between SIRTs and OS in other different subtypes of OC (Kaplan–Meier plotter).

			SIRT1		SIRT2		SIRT3		SIRT4		SIRT5		SIRT6		SIRT7	
Subtypes		**Cases**	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***
Stage	1	74	2.38 (0.75–7.54)	0.13	0.34 (0.11–1.08)	0.056	2.31 (0.62–8.55)	0.2	0.51 (0.14–1.88)	0.3	5.65 (1.13–28.18)	**0.017**	0.31 (0.1–0.96)	**0.033**	0.28 (0.09–0.88)	0.02
	2	61	1.68 (0.56–5.06)	0.35	0.64 (0.21–1.9)	0.42	0.55 (0.18–1.66)	0.28	0.38 (0.12–1.18)	0.082	0.27 (0.05–1.44)	0.1	0.3 (0.07–1.36)	0.099	0.52 (0.18–1.51)	0.22
	3	1044	1.21 (1.03–1.42)	**0.024**	0.92 (0.77–1.09)	0.33	0.75 (0.63–0.88)	**0.0005**	1.29 (1.08–1.54)	**0.005**	0.7 (0.55–0.91)	**0.0064**	0.77 (0.65–0.91)	**0.0017**	0.79 (0.66–0.94)	**0.0093**
	4	176	0.82 (0.55–1.21)	0.31	1.48 (1.02–2.14)	**0.036**	1.28 (0.88–1.87)	0.19	1.31 (0.86–2)	0.21	0.69 (0.36–1.34)	0.27	0.66 (0.43–1)	**0.046**	1.27 (0.84–1.92)	0.26
Grade	1+2	380	1.28 (0.95–1.73)	0.1	1.15 (0.85–1.56)	0.37	0.59 (0.44–0.79)	**0.0004**	0.8 (0.58–1.1)	0.17	0.61 (0.39–0.94)	**0.024**	0.62 (0.46–0.83)	**0.0011**	0.76 (0.57–1.02)	0.064
	3	1015	1.18 (0.99–1.41)	0.072	1.1 (0.92–1.33)	0.3	0.84 (0.7–1)	0.052	1.32 (1.1–1.58)	**0.003**	0.81 (0.61–1.08)	0.16	0.82 (0.69–0.97)	**0.018**	0.73 (0.61–0.88)	**0.0008**
TP53	Mutated	506	1.21 (0.95–1.53)	0.13	1.55 (1.22–1.97)	**0.0003**	1.42 (1.1–1.84)	**0.0072**	1.19 (0.94–1.49)	0.14	0.5 (0.32–0.77)	**0.0015**	1.17 (0.92–1.49)	0.21	1.19 (0.94–1.5)	0.15
	WT	94	1.53 (0.85–2.76)	0.15	0.64 (0.37–1.12)	0.12	0.54 (0.29–0.99)	**0.043**	1.69 (0.95–3.02)	0.072	0.51 (0.16–1.64)	0.25	0.67 (0.38–1.18)	0.17	1.3 (0.72–2.34)	0.38

The bold font indicates the difference was significant statistically. OC, ovarian cancer; OS, overall survival; WT, wild type.

TABLE 4 | The relationship between sirtuins and PFS in other different subtypes of OC (Kaplan–Meier plotter).

			SIRT1		SIRT2		SIRT3		SIRT4		SIRT5		SIRT6		SIRT7	
Subtypes		**Cases**	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***	**HR(95% CI)**	***P***
Stage	1	96	3.74 (1.17–11.99)	**0.018**	3.17 (1.06–9.52)	**0.03**	4.26 (1.33–13.62)	**0.0077**	2 (0.67–5.97)	0.21	3.11 (0.89–10.8)	0.06	0.43 (0.14–1.31)	0.13	2.85 (0.95–8.51)	**0.0498**
	2	67	2.04 (0.99–4.21)	**0.049**	0.74 (0.36–1.52)	0.41	0.52 (0.24–1.14)	0.096	0.63 (0.31–1.3)	0.21	2.64 (1.03–6.76)	**0.036**	0.6 (0.3–1.21)	0.15	0.63 (0.28–1.41)	0.26
	3	919	0.88 (0.75–1.04)	0.13	1.42 (1.21–1.66)	**1.5e−05**	1.15 (0.99–1.34)	0.069	1.27 (1.09–1.48)	**0.0025**	1.13 (0.89–1.43)	0.3	1.27 (1.08–1.51)	**0.0048**	1.19 (1–1.43)	0.056
	4	162	0.88 (0.59–1.3)	0.52	1.88 (1.27–2.8)	**0.0015**	1.77 (1.21–2.59)	**0.0028**	0.73 (0.5–1.08)	0.11	1.68 (0.98–2.86)	0.056	0.71 (0.48–1.06)	0.096	0.8 (0.55–1.16)	0.24
Grade	1+2	293	1.31 (0.99–1.74)	0.061	1.45 (1.08–1.94)	**0.012**	0.7 (0.51–0.95)	**0.023**	1.57 (1.16–2.12)	**0.0032**	0.73 (0.48–1.09)	0.12	0.78 (0.59–1.04)	0.085	0.79 (0.58–1.1)	0.16
	3	837	0.85 (0.71–1.01)	0.064	1.31 (1.09–1.57)	**0.0039**	1.24 (1.05–1.46)	**0.012**	1.31 (1.11–1.55)	**0.0015**	1.29 (0.99–1.68)	0.063	0.88 (0.73–1.06)	0.17	0.89 (0.74–1.07)	0.22
TP53	mutated	483	1.33 (1.05–1.68)	**0.018**	1.65 (1.32–2.06)	**1.1e−05**	1.53 (1.21–1.94)	**0.00042**	1.17 (0.94–1.47)	0.16	0.75 (0.5–1.11)	0.15	1.43 (1.12–1.82)	0.0037	1.36 (1.05–1.76)	**0.019**
	WT	84	1.84 (1.07–3.18)	**0.026**	1.91 (1.01–3.63)	**0.043**	0.67 (0.36–1.23)	0.19	1.66 (0.97–2.86)	0.063	1.78 (0.65–4.86)	0.26	1.51 (0.86–2.66)	0.15	1.44 (0.85–2.45)	0.17

The bold font indicates the difference was significant statistically. OC, ovarian cancer; PFS, progression-free survival; WT, wild type.

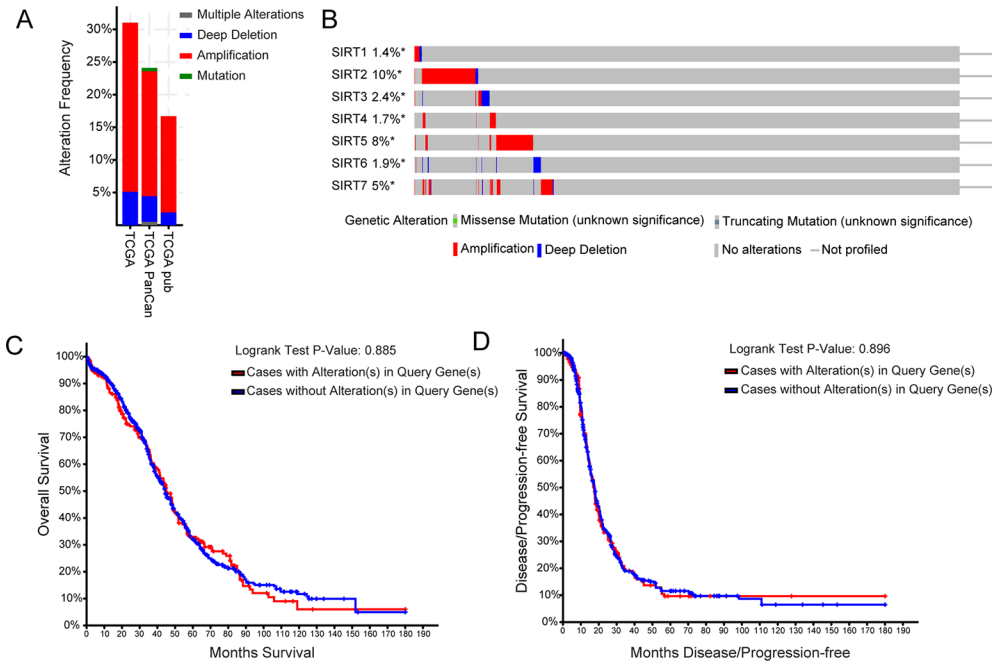
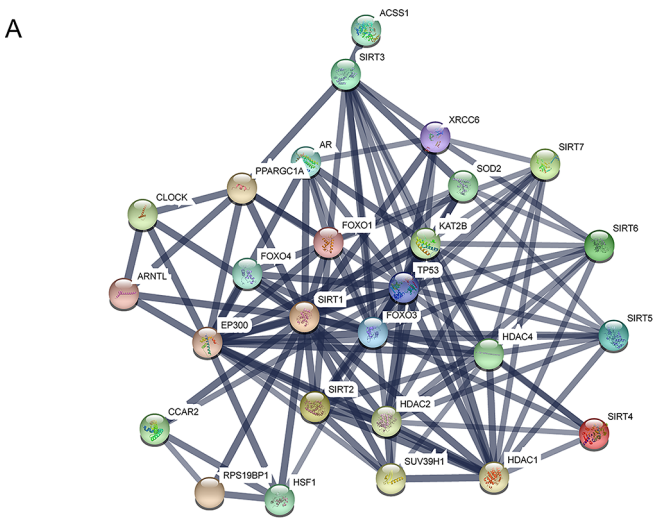


FIGURE 5 | The genetic alteration analysis of SIRTs in patients with OC (cBioPortal). **(A)** Summary of alteration in SIRTs. **(B)** OncoPrint tab summary of alteration on a query of SIRTs. Kaplan–Meier plots comparing **(C)** overall survival (OS) and **(D)** progression-free survival (PFS) in cases with/without *SIRTs* gene alterations.



B

	SIRT1	SIRT2	SIRT3	SIRT4	SIRT5	SIRT6	SIRT7
SIRT1	1	0.073	0.39	0.27	0.27	0.09	0.21
SIRT2	0.073	1	0.19	0.16	0.084	0.13	0.22
SIRT3	0.39	0.19	1	0.33	0.27	0.2	0.27
SIRT4	0.27	0.16	0.33	1	0.15	0.12	0.14
SIRT5	0.27	0.084	0.27	0.15	1	0.11	0.21
SIRT6	0.09	0.13	0.2	0.12	0.11	1	0.33
SIRT7	0.21	0.22	0.27	0.14	0.21	0.33	1

1
0.5
0

FIGURE 6 | The protein–protein interaction (PPI) of SIRTs. **(A)** The network of 7 SIRT members and 20 proteins that significantly interacted with SIRTs (String). **(B)** The Pearson correlation coefficients between SIRTs (GEPIA).

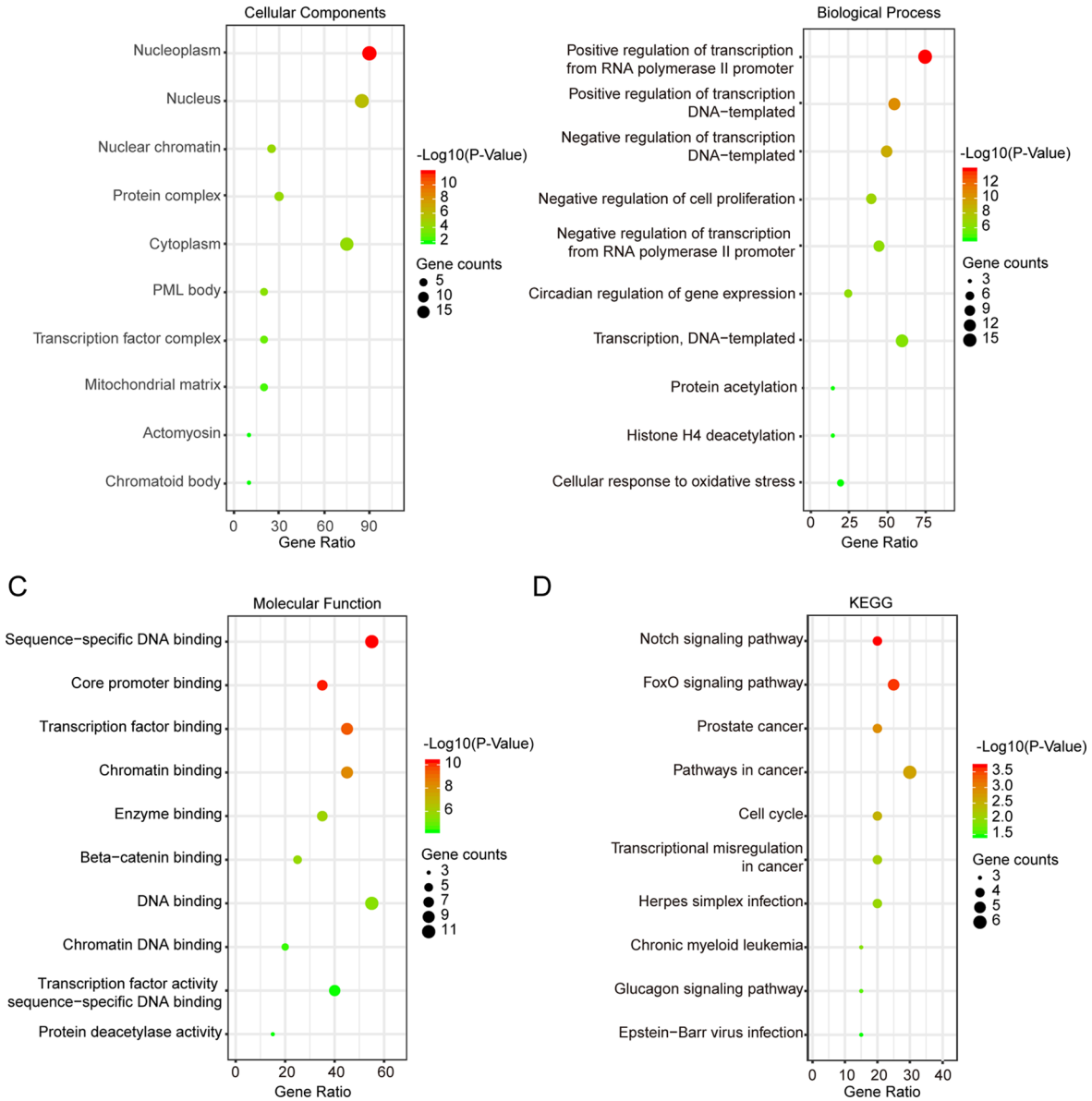


FIGURE 7 | Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of SIRTs and their interactors (DAVID). GO enrichment analysis of target genes based on following three aspects: **(A)** cellular component, **(B)** biological process, and **(C)** molecular function. **(D)** KEGG pathway enrichment analysis of target genes.

that SIRTs and their 20 interactors were mainly correlated with cancer-related pathways such as the Notch and FOXO pathways.

Despite the numerous findings, there are some limitations to this study. First, this was a bioinformatics analysis mainly based on transcriptional data, whereas proteins are the primary mediators of the various functions. Moreover, although SIRTs showed distinct prognostic values in OC, the multivariable analyses of molecules such as breast cancer type 1, human epididymis protein 4, and cancer antigen 125 are needed for further identification. Thus, the utility of SIRT expression as independent prognostic indicators in OC is yet to be further confirmed. Finally, since all the data were obtained from different databases with inevitable background heterogeneity, our results may contain some inconsistency. To address these issues, we are planning to perform well designed studies to verify these findings in the near future.

In conclusion, the mRNA expression patterns, prognostic values, genetic alterations, and PPI networks of SIRTs in OC patients were investigated. This comprehensive bioinformatics analysis revealed that SIRT1–4, 6, and 7 may be new prognostic biomarkers, and SIRT5 may be a potential target for precision therapy for patients with OC. However, further studies are needed

to confirm this notion. Finally, these findings would contribute to a better understanding of the distinct roles of SIRTs in OC.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: www.oncomine.org, http://gepia.cancer-pku.cn/, http://kmplot.com/analysis/, https://www.cbioportal.org/, https://string-db.org/, https://david.ncifcrf.gov/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Research Ethics Committee of China Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to study design, data analysis, drafting, or revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 81672964, 81874214, and 81702269).

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.cn) for English-language editing.

REFERENCES

Aljada, A., Saleh, A. M., and Al Suwaidan, S. (2014). Modulation of insulin/IGFs pathways by sirtuin-7 inhibition in drug-induced chemoreistance. *Diagn. Pathol.* 9, 1–9. doi: 10.1186/1746-1596-9-94

Aljada, A., Saleh, A. M., Alkathiri, M., Shamsa, H. B., Al-Bawab, A., and Nasr, A. (2015). Altered sirtuin 7 expression is associated with early stage breast cancer. *Breast Cancer* 9, 3–8. doi: 10.4137/BCBCR.S23156

Bae, J. S., Noh, S. J., Kim, K. M., Park, S.-H., Hussein, U. K., Park, H. S., et al. (2018). SIRT6 is involved in the progression of ovarian carcinomas *via* β-catenin-mediated epithelial to mesenchymal transition. *Front. Oncol.* 8, 1–15. doi: 10.3389/fonc.2018.00538

Barber, M. F., Michishita-kioi, E., Xi, Y., Tasselli, L., Kioi, M., Moqtaderi, Z., et al. (2013). SIRT7 links H3K18 deacetylation to maintenance of oncogenic transformation. *Nature* 487, 114–118. doi: 10.1038/nature11043

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., et al. (2011a). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Bell, E. L., Emerling, B. M., Ricoult, S. J. H., and Guarente, L. (2011b). SirT3 suppresses hypoxia inducible factor 1alpha and tumor growth by inhibiting mitochondrial ROS production. *Oncogene* 30, 2986–2996. doi: 10.1038/onc.2011.37

Bonome, T., Levine, D. A., Shih, J., Randonovich, M., Pise-Masison, C. A., Bogomolniy, F., et al. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res.* 68, 5478–5486. doi: 10.1158/0008-5472.CAN-07-6595

Chalkiadaki, A., and Guarente, L. (2015). The multifaceted functions of sirtuins in cancer. *Nat. Rev. Cancer* 15, 608–624. doi: 10.1038/nrc3985

Chen, W., Yuan, H., and Su, L. (2013). The emerging and diverse roles of sirtuins in cancer: a clinical perspective. *Onco Targets Ther.* 6, 1399–1416. doi: 10.2147/OTT.S37750

Csibi, A., Fendt, S.-M., Li, C., Poulogiannis, G., Choo, A. Y., Chapski, D. J., et al. (2013). The mTORC1 pathway stimulates glutamine metabolism and cell proliferation by repressing SIRT4. *Cell* 153, 840–854. doi:10.1016/j.cell.2013.04.023.

Dong, X. C., Jing, L. M., Wang, W. X., and Gao, Y. X. (2016). Down-regulation of SIRT3 promotes ovarian carcinoma metastasis. *Biochem. Biophys. Res. Commun.* 475, 245–250. doi: 10.1016/j.bbrc.2016.05.098

Dryden, S. C., Nahhas, F. A., Nowak, J. E., Goustin, A.-S., and Tainsky, M. A. (2003). Role for human SIRT2 NAD-dependent deacetylase activity in control of mitotic exit in the cell cycle. *Mol. Cell. Biol.* 23, 3173–3185. doi: 10.1128/MCB.23.9.3173-3185.2003

Du, J., Zhou, Y., Su, X., Yu, J. J., Khan, S., Jiang, H., et al. (2011). Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* 334, 806–809. doi: 10.1126/science.1207861

Du, Y., Wu, J., Zhang, H., Li, S., and Sun, H. (2017). Reduced expression of SIRT2 in serous ovarian carcinoma promotes cell proliferation through disinhibition of CDK4 expression. *Mol. Med. Rep.* 15, 1638–1646. doi: 10.3892/mmr.2017.6183

Fernandez-marcos, P. J., and Serrano, M. (2013). Europe PMC Funders Group Sirt4: the glutamine gatekeeper. *Cancer Cell* 23, 427–428. doi: 10.1016/j.ccr.2013.04.003

Finley, L. W. S., Carracedo, A., Lee, J., Souza, A., Egia, A., Zhang, J., et al. (2011). SIRT3 opposes reprogramming of cancer cell metabolism through HIF1alpha destabilization. *Cancer Cell* 19, 416–428. doi: 10.1016/j.ccr.2011.02.014

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, pl1. doi: 10.1126/scisignal.2004088

Győrffy, B., Lánczky, A., and Szállási, Z. (2012). Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer* 19, 197–208. doi: 10.1530/ERC-11-0329

Hendrix, N. D., Wu, R., Kuick, R., Schwartz, D. R., Fearon, E. R., and Cho, K. R. (2006). Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. *Cancer Res.* 66, 1354–1362. doi: 10.1158/0008-5472.CAN-05-3694

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Hubbi, M. E., Hu, H., Kshitiz, Gilkes, D. M., and Semenza, G. L. (2013). Sirtuin-7 inhibits the activity of hypoxia-inducible factors. *J. Biol. Chem.* 288, 20768–20775. doi: 10.1074/jbc.M113.476903

Hyde, A. R., Taylor, A. S., and Batishchev, O. V. (2018). Impaired DNA damage response, genome instability, and tumorigenesis in SIRT1 mutant mice. *IEEE Trans. Plasma Sci.* 46, 395–405. doi: 10.1109/TPS.2018.2790362

Jang, K. Y., Kim, K. S., Hwang, S. H., Kwon, K. S., Kim, K. R., Park, H. S., et al. (2009). Expression and prognostic significance of SIRT1 in ovarian epithelial tumours. *Pathology* 41, 366–371. doi: 10.1080/00313020902884451

Kim, H.-S., Vassilopoulos, A., Wang, R.-H., Lahusen, T., Xiao, Z., Xu, X., et al. (2011). SIRT2 maintains genome integrity and suppresses tumorigenesis through regulating APC/C activity. *Cancer Cell* 20, 487–499. doi: 10.1016/j.ccr.2011.09.004

Li, J., Yue, H., Yu, H., Lu, X., and Xue, X. (2019). Development and validation of SIRT3-related nomogram predictive of overall survival in patients with serous ovarian cancer. *J. Ovarian Res.* 12, 1–9. doi: 10.1186/s13048-019-0524-2

Lu, K. H., Patterson, A. P., Wang, L., Marquez, R. T., Atkinson, E. N., Baggerly, K. A., et al. (2004). Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis. *Clin. Cancer Res.* 10, 3291–3300. doi: 10.1158/1078-0432.CCR-03-0409

Mvunta, D. H., Miyamoto, T., Asaka, R., Yamada, Y., Ando, H., Higuchi, S., et al. (2017). Overexpression of SIRT1 is associated with poor outcomes in patients with ovarian carcinoma. *Appl. Immunohistochem. Mol. Morphol.* 25, 415–421. doi: 10.1097/PAI.0000000000000316

Nakagawa, T., Lomb, D. J., Haigis, M. C., and Guarente, L. (2009). Regulates the urea cycle. *Cell* 137, 560–570. doi: 10.1016/j.cell.2009.02.026

O'Callaghan, C., and Vassilopoulos, A. (2017). Sirtuins at the crossroads of stemness, aging, and cancer. *Aging Cell* 16, 1208–1218. doi: 10.1111/acel.12685

Osborne, B., Bentley, N. L., Montgomery, M. K., and Turner, N. (2016). The role of mitochondrial sirtuins in health and disease. *Free Radic. Biol. Med.* 100, 164–174. doi: 10.1016/j.freeradbiomed.2016.04.197

Qin, J., Liu, Y., Lu, Y., Liu, M., Li, M., Li, J., et al. (2017). Hypoxia-inducible factor 1 alpha promotes cancer stem cells-like properties in human ovarian cancer cells by upregulating SIRT1 expression. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-09244-8

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia (New York, N.Y.)* 6, 1–6. doi: 10.1016/S1476-5586(04)80047-2

Serrano, L., Martinez-Redondo, P., Marazuela-Duque, A., Vazquez, B. N., Dooley, S. J., Voigt, P., et al. (2013). The tumor suppressor SirT2 regulates cell cycle progression and genome stability by modulating the mitotic deposition of H4K20 methylation. *Genes Dev.* 27, 639–653. doi: 10.1101/gad.211342.112

Shuang, T., Wang, M., Zhou, Y., and Shi, C. (2015). Over-expression of Sirt1 contributes to chemoresistance and indicates poor prognosis in serous epithelial ovarian cancer (EOC). *Med. Oncol.* 32, 1–7. doi: 10.1007/s12032-015-0706-8

Tan, M., Peng, C., Anderson, K. A., Chhoy, P., Xie, Z., Dai, L., et al. (2014). Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. *Cell Metab.* 19, 605–617. doi: 10.1016/j.cmet.2014.03.014

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

Teng, C., and Zheng, H. (2017). Low expression of microRNA-1908 predicts a poor prognosis for patients with ovarian cancer. *Oncol. Lett.* 14, 4277–4281. doi: 10.3892/ol.2017.6714

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Wang, H. L., Lu, R. Q., Xie, S. H., Zheng, H., Wen, X. M., Gao, X., et al. (2015). SIRT7 exhibits oncogenic potential in human ovarian cancer cells. *Asian Pac. J. Cancer Prev.* 16, 3573–3577. doi: 10.7314/APJCP.2015.16.8.3573

Xiang, X. Y., Kang, J. S., Yang, X. C., Su, J., Wu, Y., Yan, X. Y., et al. (2016). SIRT3 participates in glucose metabolism interruption and apoptosis induced by BH3 mimetic S1 in ovarian cancer cells. *Int. J. Oncol.* 49, 773–784. doi: 10.3892/ijo.2016.3552

Yoshihara, K., Tajima, A., Komata, D., Yamamoto, T., Kodama, S., Fujiwara, H., et al. (2009). Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. *Cancer Sci.* 100, 1421–1428. doi: 10.1111/j.1349-7006.2009.01204.x

Zhang, J., Yin, X. J., Xu, C. J., Ning, Y. X., Chen, M., Zhang, H., et al. (2015). The histone deacetylase SIRT6 inhibits ovarian cancer cell proliferation *via* down-regulation of Notch 3 expression. *Eur. Rev. Med. Pharmacol. Sci.* 19, 818–824.

Zhou, Q., Hou, C. N., Yang, H. J., He, Z., and Zuo, M. Z. (2018). Distinct expression and prognostic value of members of the epidermal growth factor receptor family in ovarian cancer. *Cancer Manag. Res.* 10, 6937–6948. doi: 10.2147/CMAR.S183769

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 September 2019
doi: 10.3389/fgene.2019.00831

Prognostic Roles of Central Carbon Metabolism–Associated Genes in Patients With Low-Grade Glioma

Li Wang[1†], Meng Guo[2†], Kai Wang[1] and Lei Zhang[1*]*

[1] Department of Neurosurgery, Xijing Hospital, Fourth Military Medical University, Xi'an, China, [2] Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Xi'an, China

Purpose: Metabolic alterations are crucial for tumor progression and response to therapy. The comprehensive model of combined central carbon metabolism–associated genes that contribute to the outcomes of glioma and astrocytoma is not well understood.

Method: We studied the profiles of 63 genes involved in central carbon metabolism in 514 relatively low-grade glioma patients. The different distributions of gene expression in gliomas and astrocytoma were identified. The differential gene expression between each cohort and the correlations with prognosis were detected. Finally, we built a tentative model to detect the prognostic roles of carbon metabolism–associated genes in astrocytoma.

Result: Two primary clusters and four subclusters with significantly different overall survival were identified in low-grade glioma. The differences of histological diagnoses, grade, tumor site, and age were detected between each cluster. Comparing with other histological types, patients with astrocytoma exhibited the worst prognosis. Between astrocytoma patients with poor and favorable prognoses, expression profiles of 11 genes were significantly discrepant. We detected that 18 genes were respectively correlated with overall survival in astrocytoma; moreover, four genes (*RAF1*, *AKT3*, *IDH1*, and *FGFR1*) were detected as dependent variables for the prediction of the survival status of astrocytoma patients and were capable to predict the survival.

Conclusion: Central carbon metabolism–associated genes are differentially expressed in all patients with glioma and histological subtype astrocytoma. The gene expression profile is significantly associated with clinical manifestations. These results suggested that both the multigene expression patterns and individual central carbon metabolism–associated genes were potentially capable to predict the prognosis of patients with low-grade glioma.

Keywords: low-grade glioma, astrocytoma, prognosis, metabolism, gene expression

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Yang An,
Henan University, China
Haiwei Mou,
Cold Spring Harbor Laboratory, United States

***Correspondence:**
Kai Wang
wkslashking@163.com
Lei Zhang
zhangleiafmmu@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *08 May 2019*
Accepted: *12 August 2019*
Published: *18 September 2019*

Citation:
Wang L, Guo M, Wang K and Zhang L (2019) Prognostic Roles of Central Carbon Metabolism–Associated Genes in Patients With Low-Grade Glioma. Front. Genet. 10:831. doi: 10.3389/fgene.2019.00831

INTRODUCTION

Diffuse low-grade gliomas are the most common primary malignancies in adults and include astrocytomas, oligodendrogliomas, and oligoastrocytomas (Brat et al., 2015). Different histological subtypes of glioma were undistinguishable; however, large differences in clinical behavior and response to therapy suggest that difference among the histological types is crucial (Smith et al., 2000). Even within each subtype, there are large differences in clinical performance among individual patients. Surgery resection is the primary therapeutic method for low-grade gliomas,

but the outcomes are less than satisfactory because of the highly infiltrative nature of glioma, and the presence of residual tumor tissue results in recurrence and malignant progression (Dixit and Raizer, 2017). The prognosis of patients with relatively low-grade glioma varies widely, with some patients living for more than 5 years, while others survive less than 1 year (Bush and Chang, 2016). A more precise method of predicting the outcomes of relatively low-grade glioma is urgently needed to be developed.

Metabolic reprogramming is a central hallmark of cancer. Dysregulation of metabolism-related genes leads to cellular transformation and tumor progression. Warburg (1956) revealed differences in the central metabolic pathways in solid tumors and noted that cancer cells require a large amount of glucose to maintain a high rate of glycolysis even in the presence of adequate oxygen and that they convert a majority of that glucose into lactic acid (the Warburg effect). More recently, it has been recognized that the "Warburg effect" contains a similarly increased utilization of glutamine (Reitzer et al., 1979). Previous studies have detected some variations in the genes, such as *IDH1/2*, *GLUT1*, and *GLUT3*, involved in tumor metabolism in gliomas (Yan et al., 2009; Verhaak et al., 2010; Labak et al., 2016). High-throughput sequencing has substantially advanced the understanding of the metabolic changes in low-grade gliomas by detecting changes in metabolism-associated genes (Brennan et al., 2013). Profiling holistic gene expression not only facilitates the investigation of subgroups with low-grade glioma but also enables the identification of the predictors of overall survival (OS) (Chen et al., 2016). Which pattern of expression of metabolism-associated genes in tumor tissue contributes to glioma is not well understood. The Cancer Genome Atlas (TCGA) provided a standardized gene expression dataset for the study of the expression pattern of metabolism-related genes, which enables the investigation of correlations between clinical manifestations and carbon metabolism–associated genes in glioma (The Cancer Genome Atlas Research Network et al., 2008; Sanborn et al., 2013).

In this study, we investigated the expression patterns of central carbon metabolism–associated genes in adult patients with diffuse low-grade glioma, including astrocytoma, oligodendroglioma, and oligoastrocytoma. Moreover, we respectively detected the prognostic roles of individual gene and the multiple-gene combination. These results will facilitate an integral understanding of the metabolic alterations in glioma and provide a novel perspective to manage and treat this lethal cancer.

METHODS

Samples and Database

We obtained transcriptome data and the corresponding clinical data of 514 relatively low-grade glioma patients from TCGA from the cBioPortal for Cancer Genomics (http://cbioportal.org) (Gao et al., 2013). We filtered the data based on whether the mRNA *z*-score data, histological diagnosis, and OS data were comprehensive. Collectively, the studied dataset included 194 astrocytoma samples, 130 oligoastrocytoma samples, and 190 oligodendroglioma samples.

Central carbon metabolism–related genes in the cancer-associated gene panel (hsa05230) were derived from the KEGG pathway database (http://www.kegg.jp/kegg/), as previously described (Kanehisa et al., 2017). In total, 65 central carbon metabolism–associated genes were listed; however, transcriptome information was missing for *MYC* and *HKDC1*, and the remaining 63 candidate genes were included after filtration. The gene expression levels were calculated from the mRNA *z* scores and compared to the expression distribution of each gene from tumors that were diploid for the genes in 514 patients with glioma (RNA-Seq V2 RSEM), based on TCGA data.

Bioinformatics

A cluster analysis of the 63 genes expressed in each histological type was used to distinguish samples based on gene expression patterns. Samples with different gene expression patterns were identified from the whole dataset. The transcriptional levels were shown as mRNA *z* scores and clustered using the hierarchical clustering algorithm in the Gene Cluster 3.0 program (De Hoon et al., 2004). The cluster heat map and pattern according to tumor stage were generated with the Java Treeview program (Saldanha, 2004).

Prognostic Implication Analyses

To investigate the prognostic role of the cancer metabolism–associated genes, we used GraphPad Prism 6 for Windows (GraphPad Software, Inc., CA, US; version 6.01, 2012) to perform comparisons of the overall survivals in different clusters. Additionally, an analysis of the difference in OS between the cohorts with low and high expression levels of differentially expressed genes was conducted with GraphPad Prism 6.

Statistical Analysis

Survival curves were plotted according to the Kaplan–Meier method and compared using the log-rank test in GraphPad Prism 6. Associations between clinical characteristics and the variables used to determine the clusters of patients were analyzed by Fisher exact test and the Pearson/Spearman correlation. Differences in gene expression levels between clusters were analyzed by analysis of variance. Correlations between variable were determined by regression analyses. All tests were performed with SPSS 19.0 (IBM, Inc., NY, US). $P < 0.05$ was considered statistically significant.

RESULTS

Expression Profile of Central Carbon Metabolism–Associated Genes in Diffuse Gliomas

To investigate central carbon metabolism programming in diffuse gliomas, we first examined the transcriptional distributions of carbon metabolism–associated genes. In total, 63 genes that have been widely reported to be key players in metabolic reprogramming

were included (Soga, 2013). The patients with diffuse glioma were sorted by differences in the gene expression according to the RNA-Seq data. Following filtration, 514 patients with survival data were included in the cluster analysis (**Figure 1A**). The preliminary analysis showed that there were two clusters, and strikingly, the 101 patients in cluster 1 had much worse prognoses than the patients in cluster 2 (OS of 48.65 vs 105.12 months, $P < 0.0001$) (**Figure 1B**, **Table 1**). Between the two clusters, there was a significant difference in the expression levels ($P < 0.05$) of 49 genes (**Figure 1C**). A comparison of the clinical characteristics of clusters 1 and 2

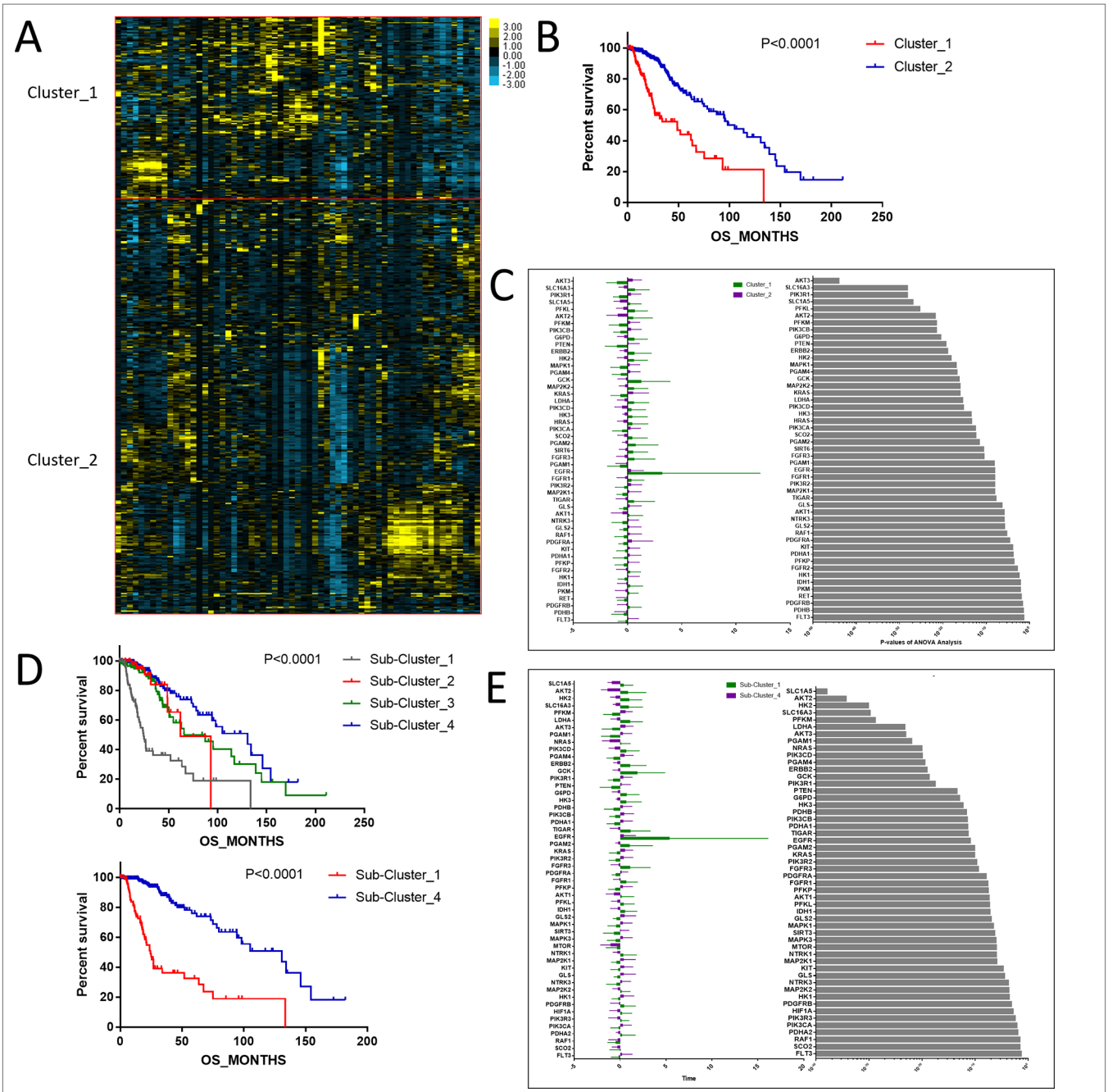


FIGURE 1 | The expression profile of central carbon metabolism–associated genes in glioma patients. **(A)** In total, 514 patients were primarily divided into two clusters. The expression values of 63 genes corresponding to the individual patient were arrayed in the columns according to the expression affinity. Patients with similar gene expression patterns were clustered and grouped using the hierarchical clustering algorithm and arrayed in the rows. **(B)** The patients in cluster 1 had much worse prognoses than the patients in cluster 2, of which overall survival (OS) was 48.65 months compared to 105.12 months. **(C)** There were 49 genes that showed a significant difference in the expression levels between the two clusters $P < 0.05$. **(D)** The studied cohort was further subdivided into four subclusters, among which the subcluster 1 was with the worst OS and subcluster 4 showed the most favorable outcome. **(E)** The differential expression analysis revealed that 52 metabolism-associated genes were significantly different between subcluster 1 and subcluster 4.

TABLE 1 | Overall survival differences of each cluster.

	Cluster		Significance (P)
	1	2	
Number	158	356	<0.0001
Median survival	48.65	105.12	

showed that the parameters of histological diagnoses, tumor grade, tumor site, and age were vastly different between the two clusters ($P < 0.001$), as shown in **Table 2**.

To detect the subtler differences among further stratified cohorts, we subdivided the 514 patients into four subclusters based on the expression of the 63 genes. We detected that extreme differences in prognosis were shown among those subclusters ($P < 0.0001$) (**Figure 1D**, **Table 3**). Subcluster 1 was associated with the worst OS, while subcluster 4 showed a much favorable outcome than other subclusters (**Figure 1D**). Comparison of the gene expression variations revealed that the expression levels of 52 metabolism-associated genes were significantly different between the two prognostic-discrepant cohorts (**Figure 1E**). Additionally, we compared the clinical characteristics of the subclusters. Similar with the previous result, clear significant differences were found with regard to the parameters of histological diagnoses, tumor grade, tumor site, and age ($P < 0.001$) (**Table 4**).

Variations in Metabolism-Associated Gene Expression Levels in Different Histological Types

According to the binary comparisons among different gene expression cohorts, histological type was revealed as the variable associated with the largest differences in gene expression. We compared the OS of patients with astrocytoma, oligoastrocytoma, and oligodendroglioma, and significant differences were detected (**Figure 2A**). The median survival times of patients with astrocytoma (66.12 months), oligoastrocytoma 5.12 months), and oligodendroglioma (95.5 months) were markedly distinguishing ($P = 0.0084$). Further analysis of differences in gene expression demonstrated that 45 metabolism-associated genes were differentially expressed among the histological types (**Figure 2B**, **Supplementary Table 1**).

Differences in the Expression Levels of Metabolism-Associated Genes in Astrocytoma

The above results showed that among the histological types of glioma, astrocytoma showed the worst prognosis. To study the expression profiles of metabolism-associated genes in the poor-prognosis histological types, we grouped the patients with astrocytoma according to the metabolism-associated

TABLE 2 | Characteristics of glioma patients in clustered groups 1 and 2.

Clinical features		Cluster		Total	P
		1	2		
Histological diagnosis	Astrocytoma	92	102	194	2.96E–10***
	Oligoastrocytoma	34	96	130	
	Oligodendroglioma	32	158	190	
Grade	Unknown	0	1	1	<0.001***
	G2	52	196	248	
	G3	106	159	265	
Age	≤41	53	215	268	1.37E–8***
	>41	105	141	246	
Tumor site	Unknown	0	1	1	<0.001***
	Posterior fossa, brain stem	1	0	1	
	Posterior fossa, cerebellum	2	0	2	
	Supratentorial, frontal lobe	73	229	302	
	Supratentorial, not otherwise specified	7	1	8	
	Supratentorial, occipital lobe	0	8	8	
	Supratentorial, parietal lobe	18	28	46	
	Supratentorial, temporal lobe	57	89	146	
Laterality	Unknown	1	4	5	0.412
	Left	79	171	250	
	Midline	4	3	7	
	Right	74	178	252	
Supratentorial localization	Unknown	7	22	29	0.853
	Cerebral cortex	43	98	141	
	Deep gray	1	2	3	
	Not listed in medical Record	73	149	222	
	White matter	34	85	119	
Sex	Female	74	155	229	0.275
	Male	84	201	285	
History neoadjuvant	No	158	353	511	0.719
	Yes	0	3	3	

*** $P < 0.001$.

TABLE 3 | Overall survival differences of each subcluster

	Subcluster				Significance (*P*)
	1	2	3	4	
Number	101	57	123	233	<0.0001
Median survival	24.38	62.12	87.39	130.68	

genes transcriptional data. Among the patients, two primary clusters that showed distinguishing median survival times of 43.99 and 73.42 months were identified ($P = 0.0064$) (**Figures 3A**, **B**). Further, four subclusters were divided according to fine grouping. The comparison of the OS showed that the difference in prognosis was even more marked ($P < 0.0001$) (**Figure 3C**). Subcluster 2 had the worse prognosis (median of 24.9 months) than other subclusters (median of 67.41 months), $P < 0.0001$ (**Figure 3D**). We respectively detected the gene expression differences between cluster 1 versus cluster 2; among different subclusters and subcluster 2 versus the other subclusters, the results revealed that the expressions of 33 metabolism-associated genes were significantly varied (**Figure 3E**).

The Prognostic Role of Metabolism-Associated Genes in Astrocytoma

We uncovered that the expression pattern of metabolism-associated genes was closely related to the prognosis of patients with astrocytoma. To investigate the effect of individual metabolism-associated gene on the prognosis of astrocytoma patients, we divided the subjects into two cohorts according to the OS: poor prognosis group and good prognosis group. We further investigated the differences in the expression levels of the metabolism-associated genes. It was detected that 11 genes, namely, *FGFR1*, *ERBB2*, *PGAM4*, *PGAM1*, *G6PD*, *RET*, *AKT3*, *PTEN*, *RAF1*, *PKM*, and *LDHA*, had significantly different expression levels between patients with poor and favorable OS times (**Figure 4A**, **Supplementary Table 2**).

TABLE 4 | Characteristics of glioma patients in subdivided clusters.

Clinical features		Subcluster				Total	*P*
		1	2	3	4		
Histological diagnosis	Astrocytoma	67	25	60	42	194	5.94E–21***
	Oligoastrocytoma	19	15	40	56	130	
	Oligodendroglioma	15	17	23	135	190	
Grade	Unknown	0	0	0	1	1	3.88E–08***
	G2	20	32	60	136	248	
	G3	81	25	63	96	265	
Age	≤41	32	21	84	131	268	5.32E–8***
	>41	69	36	39	102	246	
Tumor site	Unknown	0	0	0	1	1	<0.001***
	Posterior fossa, brain stem	1	0	0	0	1	
	Posterior fossa, cerebellum	2	0	0	0	2	
	Supratentorial, frontal lobe	40	33	71	158	302	
	Supratentorial, not otherwise specified	4	3	0	1	8	
	Supratentorial, occipital lobe	0	0	3	5	8	
	Supratentorial, parietal lobe	11	7	15	13	46	
	Supratentorial, temporal lobe	43	14	34	55	146	
Laterality	Unknown	1	0	2	2	5	0.235
	Left	44	35	60	111	250	
	Midline	4	0	1	2	7	
	Right	52	22	60	118	252	
Supratentorial localization	Unknown	4	3	6	16	29	0.910
	Cerebral cortex	24	19	32	66	141	
	Deep gray	1	0	0	2	3	
	Not listed in medical record	50	23	55	94	222	
	White matter	22	12	30	55	119	
Sex*	Female	45	29	47	108	229	0.357
	Male	56	28	76	125	285	
History neoadjuvant	No	100	55	119	225	499	0.453
	Yes	1	2	4	8	15	

*** *represent P < 0.001.*

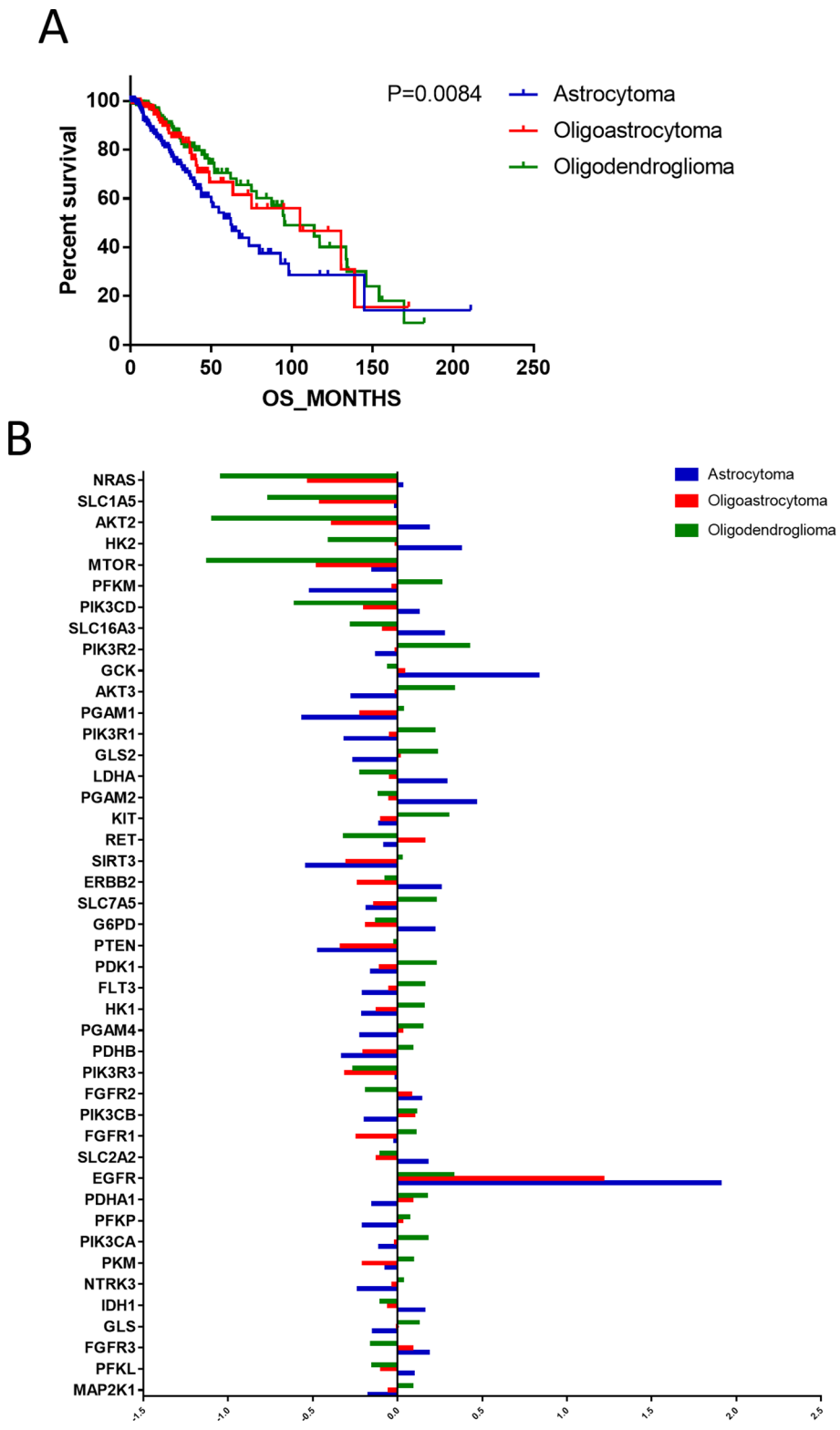


FIGURE 2 | The differences in prognoses of patients with astrocytoma, oligoastrocytoma, and oligodendroglioma were significant. **(A)** The prognoses of patients with astrocytoma, oligoastrocytoma, and oligodendroglioma were significant. **(B)** Differential expression analysis demonstrated that 45 metabolism-associated genes were discrepantly expressed among the three histological types.

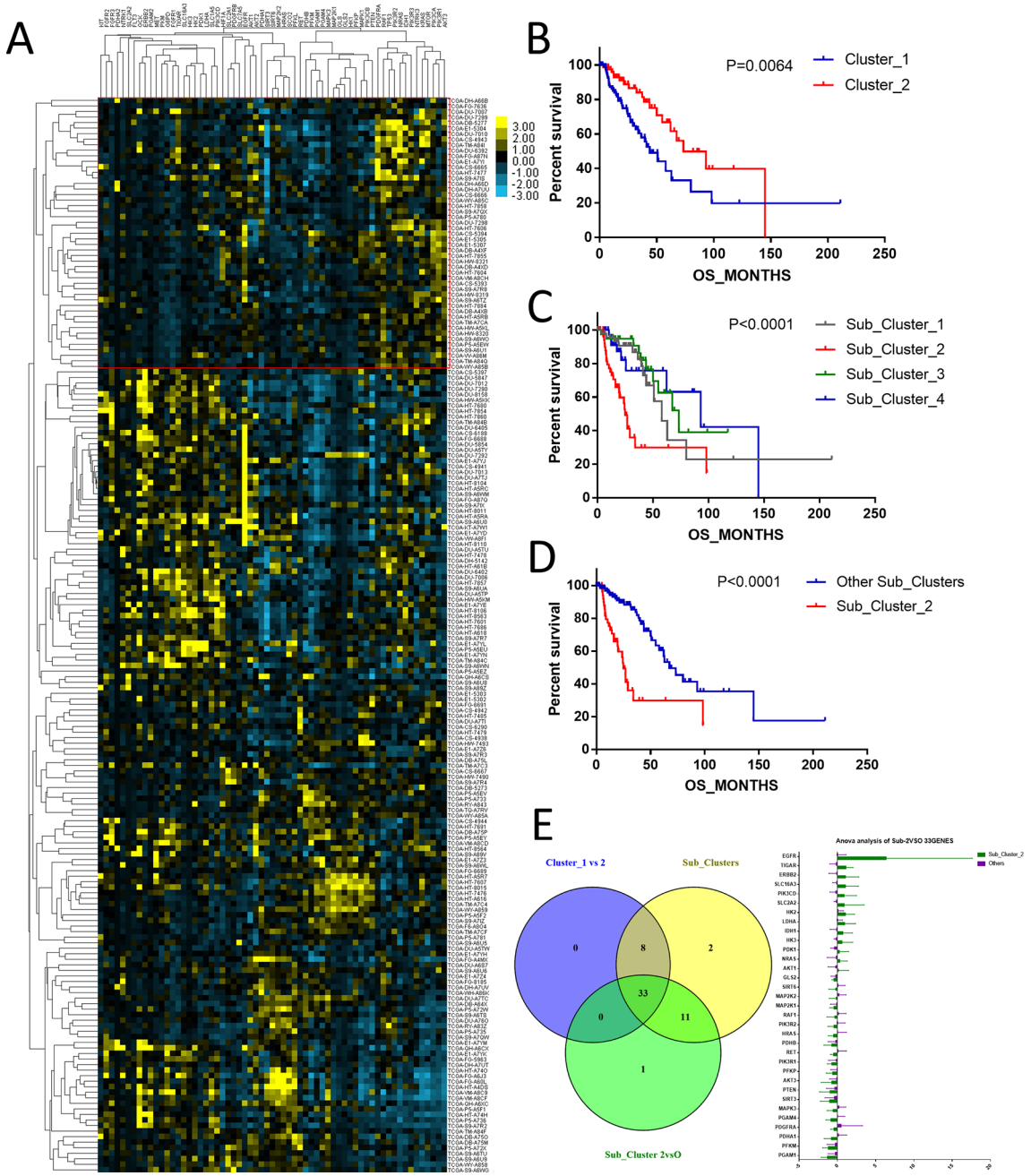


FIGURE 3 | The expression profile of central carbon metabolism–associated genes in patients with astrocytoma. **(A)** According to expression profiling, two primary clusters and additionally four subdivided clusters were identified. The comparison of median survival between two primary clusters **(B)** and five subdivided clusters **(C)** showed significant difference. In addition, patients in subcluster 2 showed the worst prognosis comparing to other patients, $P < 0.0001$ **(D)**. **(E)** Differential expression analysis demonstrated that 33 metabolism-associated genes were significantly variated in all contrast of clusters 1 and 2, subcluster 2 and the other subclusters, and five subclusters.

Additionally, we detected the pertinences between the trend of metabolism-associated gene expression differential and survival variation. According to study correlation of individual gene expression and survival, positive correlations were detected between the respective expression levels of nine genes containing *PGAM1*, *PGAM4*, *RAF1*, *PDHB*, *AKT3*, *PTEN*, *PIK3R1*, *RET*, and *MAPK3* with OS ($r > 0.2$, $P < 0.05$); on the other hand, the expression levels of nine genes containing *EGFR*, *IDH1*, *GCK*, *PKM*, *LDHA*, *G6PD*, *TIGAR*, *ERBB2*, and *FGFR1* were detected negatively correlated to survival ($r < -0.2$, $P < 0.05$) (**Table 5**). In addition to their associations with survival, the expression levels of the genes are closely correlated between the two sets (**Figure 4B**).

TABLE 5 | The correlation of overall survival (OS) of astrocytoma and expressing variation of individual gene.

OS positively correlated genes			OS negatively correlated genes		
Pearson correlation coefficient *(r)*		**Significant *(P)***	**Pearson correlation coefficient *(r)***		**Significant *(P)***
PGAM1	0.41	<0.001***	FGFR1	−0.45	<0.001***
PGAM4	0.40	0.001**	ERBB2	−0.39	0.001**
RAF1	0.36	0.003**	TIGAR	−0.37	0.002**
PDHB	0.29	0.012*	G6PD	−0.33	0.006**
AKT3	0.28	0.016*	LDHA	−0.32	0.007**
PTEN	0.27	0.020*	PKM	−0.28	0.017*
PIK3R1	0.26	0.025*	IDH1	−0.27	0.021*
RET	0.25	0.029*	GCK	−0.27	0.018*
MAPK3	0.25	0.031*	EGFR	−0.23	0.039*

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

To address the prognostic roles of those survival-related genes, we separately split the astrocytoma patients into two groups according to the single gene expression and additionally compared the prognosis between the two groups (**Figure 5**). Except to *AKT3* and *PIK3R1*, 16 genes showed a significant association with prognosis. Patients with low expression levels of *RET* and *PGAM1* were associated with a greater hazard ratio (HR) for death than that of patients with high expression levels of *RET* and *PGAM1* ($P < 0.0001$). In contrast, patients with high expression levels of *TIGAR*, *ERBB2*, *EGFR*, and *FGFR1* had a higher HR for death than that of patients with low expression levels of those genes ($P < 0.0001$) (**Table 6**).

To evaluate the effects of differences in gene expression on the prediction of the outcome of astrocytoma, we ranked the expression data of 18 genes to construct a regression model. Based on the ranking results, four genes (*RAF1*, *AKT3*, *IDH1*, and *FGFR1*) were independent predictors of the survival status of astrocytoma patients (**Supplementary Table S3**). To integrate these four genes into a single panel, multivariate Cox regression analysis was employed to obtain the coefficient. The risk score was calculated as follows: the risk score was equal to the expression of *RAF1*∗1.801 plus the expression of *AKT3*∗1.545 plus the expression of *IDH1*∗1.569 plus the expression of *FGFR1*∗1.035 (**Figure 6A**). As shown in **Figure 6B**, the area under the receiver operating characteristic curve of the four-gene panel for the prediction of the long- or short-term outcomes of astrocytoma was 0.9407, with a 95% confidence interval of 0.8864 to 0.9949 and a $P < 0.0001$.

DISCUSSION

Low-grade glioma has complicated characteristics and diverse histologic types. Although the histopathological classification of low-grade gliomas is reliable, it varies between observers and is insufficient to predict clinical outcomes (Louis et al., 2016). Recently, the molecular analysis of tumors has become a critical part of tumor classification and prognostication, and increasing evidence has suggested that defining tumor subtypes based on differences in gene expression in low-grade glioma is meaningful (Verhaak et al., 2010; Eckel-Passow et al., 2015; Louis et al., 2016). In this study, we found that metabolism-associated gene profiling was able to define two primary clusters and four subclusters of patients with low-grade glioma regardless of histologic type. Overall survival differed between the primary clusters and subclusters. We identified 44 genes with significant differences in expression levels between the groups of patients with the worst and best prognoses (**Figure S1**). Some of those genes participate in the regulation of intracellular signal transduction, and others are involved in the metabolism of glucose and other carbohydrates. In addition to the differences in gene expression, we found that the groups had significant differences in histological types, tumor grades, tumor sites, and age. The results showed the specific expression profiles of metabolism-associated genes in patients with low-grade glioma.

Astrocytomas, oligoastrocytomas, and oligodendrogliomas are the three histologic subtypes of low-grade glioma; the subtypes have always been difficult to define according to

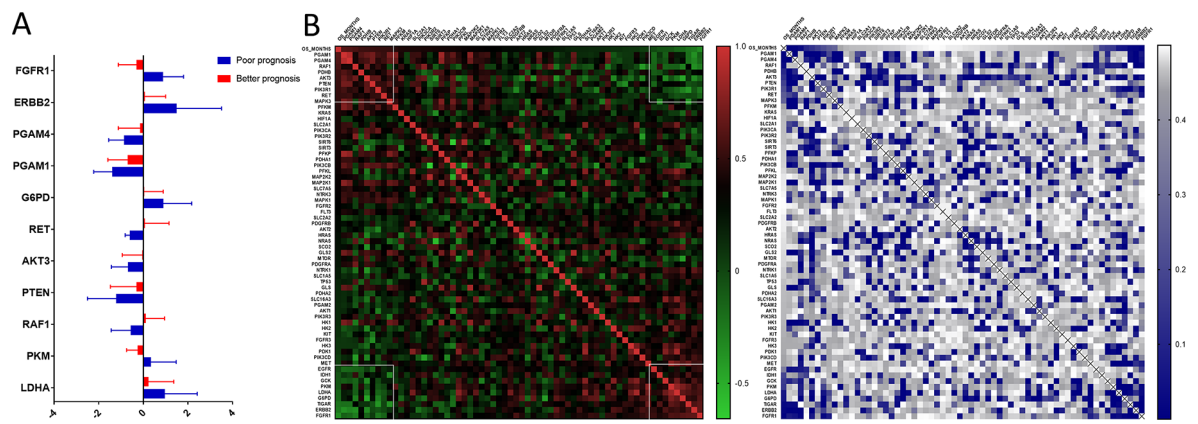


FIGURE 4 | The prognostic role of metabolism-associated genes in astrocytoma. **(A)** There are 11 differentially expressed genes, containing *FGFR1*, *ERBB2*, *PGAM4*, *PGAM1*, *G6PD*, *RET*, *AKT3*, *PTEN*, *RAF1*, *PKM*, and *LDHA*, which were detected to be significantly discrepant between patients with poor and favorable overall survival (OS) times. **(B)** According to study correlation of individual gene expression and survival, the expression levels of nine genes positively correlated with OS ($r > .2$, $P < 0.05$), and the expression levels of the other nine genes were negatively correlated with OS ($r < -0.2$, $P < 0.05$).

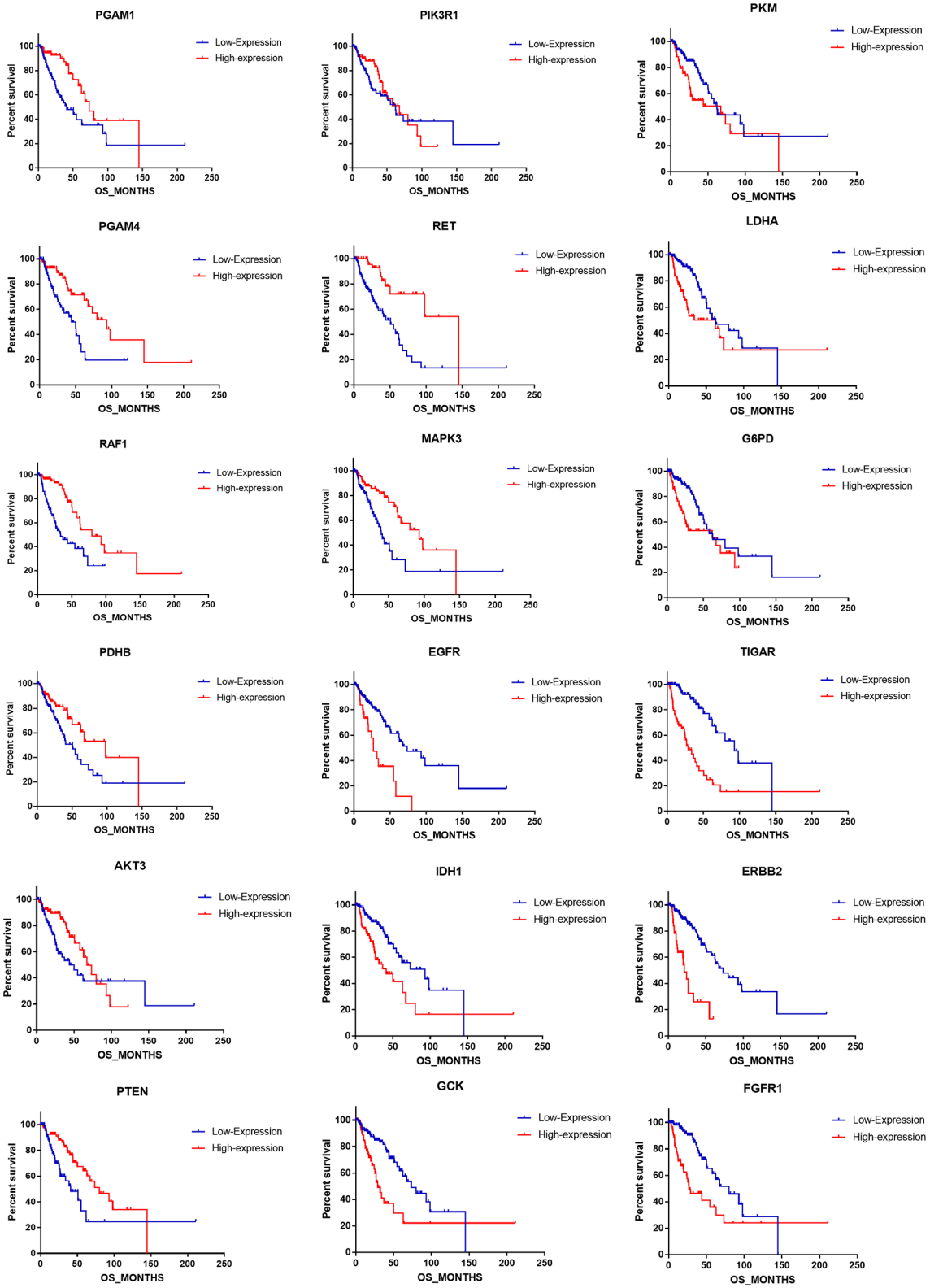


FIGURE 5 | The correlation of singular gene expression difference with the astrocytoma prognosis.

TABLE 6 | The prognostic roles of single metabolism associated gene in astrocytoma.

Gene name	Median survival (mo)		Hazard ratio	95% Confidence interval	P
	Low expression	High expression			
RAF1	33.94	79.93	0.331	1.781–5.128	<0.0001****
RET	50.82	144.94	0.277	2.145–6.069	<0.0001****
EGFR	73.42	26.91	3.007	0.155–0.715	<0.0001****
TIGAR	93.13	29.11	4.276	0.135–0.405	<0.0001****
ERBB2	73.42	21.29	4.301	0.104–0.519	<0.0001****
FGFR1	79.93	26.91	2.776	0.206–0.629	<0.0001****
GCK	73.42	29.11	2.456	0.224–0.739	0.0004***
PGAM4	43.99	93.13	0.424	1.399–3.985	0.0007***
MAPK3	39.72	93.13	0.436	1.362–3.87	0.0011**
PGAM1	41.1	73.42	0.409	1.462–4.094	0.0012**
IDH1	93.13	41.1	2.27	0.256–0.760	0.0012**
PTEN	41.1	79.93	0.475	1.235–3.589	0.0027**
LDHA	62.91	62.12	2.022	0.288–0.850	0.0060**
G6PD	62.91	62.12	1.946	0.299–0.885	0.0089**
PDHB	50.82	98.16	0.54	1.107–3.098	0.0229*
PKM	62.12	67.41	1.685	0.345–1.022	0.0441*
AKT3	43.86	67.41	0.603	0.991–2.776	0.0526
PIK3R1	62.12	67.41	0.802	0.742–2.097	0.4085

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$.

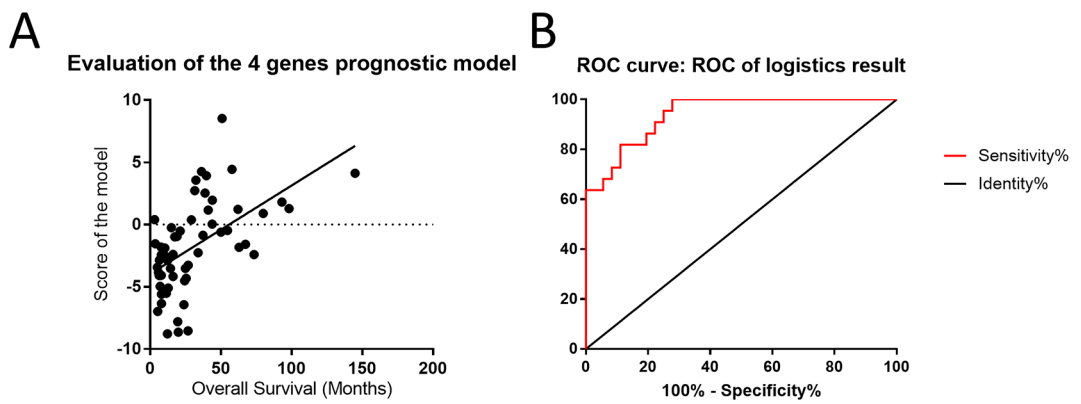


FIGURE 6 | Four-gene panel was utilized to potentially predict the outcomes of patients with astrocytoma. **(A)** The score of the four-gene model including RAF1, AKT3, IDH1, and FGFR1 was positively correlated with overall survival and showed a linearity. **(B)** The area under the receiver operating characteristic curve of the four-gene panel for the prediction of the long- or short-term outcomes of astrocytoma was 0.9407, with a 95% confidence interval of 0.8864 to 0.9949 and $P < 0.0001$.

clinical features (Louis et al., 2014). In the current dataset, patients with astrocytoma had worse prognoses than those of patients with the other two subtypes. We detected the differentially expressed genes in patients with different histological types of glioma, and 45 genes were significantly differentially expressed among the three subtypes. Moreover, 80% of those genes (35 genes) overlapped with the gene set (44 genes) that was associated with different subgroups. Specifically, we determined the expression profiles of metabolism-associated genes in astrocytomas. The results showed that 33 genes had significantly different expression levels, and those differences in expression were closely correlated with OS in patients with astrocytomas. These differences in the expression of metabolism-associated genes not only reveal metabolic differences among the histological subtypes but also suggest that there is metabolic heterogeneity within a single subtype.

In patients with astrocytomas, we identified 11 genes that varied significantly in expression between patients with poor and favorable OS. Additionally, we detected genes with expression levels that were positively and negatively associated with OS, and a correlation existed between the expression levels of these two sets of genes. According to the survival analysis, 16 genes were significantly associated with prognosis. Patients with low expression levels of *RET* and *PGAM1* and high expression levels of *TIGAR*, *ERBB2*, *EGFR*, and *FGFR1* had elevated HRs with regard to survival. The *RET* gene encodes a transmembrane receptor that is a member of the tyrosine protein kinase family of proteins. It has been reported that the mRNA levels of *RET* are elevated in astrocytoma patients with *IDH* mutations, who are known to have prolonged survival

(Zhang et al., 2018). *PGAM1* is involved in tumor cell glycolysis and biosynthesis, and this protein had elevated expression levels in high-grade astrocytomas (Liu et al., 2018). Increased expression of these two genes in astrocytomas might inhibit metabolic pathways crucial to the development and progression of tumors.

Low-grade glioma is one of the most malignant human diseases, with a very poor prognosis and scant available information about its biological properties. This study provided new information about the metabolism events affected by the identified genes with differential expression levels. We divided the patients into different subgroups according to their metabolism-associated gene expression patterns. The expression levels of those genes were strongly correlated with the prognosis of patients with astrocytoma, possibly because of their effect on the regulation of the biological behavior of the tumor. This study increases our understanding of the prognostic roles of central carbon metabolism–associated genes in patients with low-grade glioma.

DATA AVAILABILITY

The datasets analyzed for this study can be found in the cBioPortal for Cancer Genomics (http://cbioportal.org).

ETHICS STATEMENT

Data obtained from the TCGA open-access database was collected from tumors of patients who provided informed consent based on the guidelines from the TCGA Ethics, Law and Policy Group.

AUTHOR CONTRIBUTIONS

LZ designed the current study. MG collected the data and performed the statistical test. KW sorted the clinical information and interpreted the result. LW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was funded by the National Natural Science Foundation of China (no. 81702355).

ACKNOWLEDGMENTS

We also thank the Nature Research Editing Service for English language editing (certificate verification key: 32CA-EFCA-2E6F-4A78-682P).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00831/full#supplementary-material.

REFERENCES

Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., Cooper, L. A., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372, 2481–2498. doi: 10.1056/NEJMoa1402121

Brennan, C. W., Verhaak, R. G., Mckenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Bush, N.A.O., and Chang, S. (2016). Treatment strategies for low-grade glioma in adults. *J. Oncol. Pract.* 12, 1235–1241. doi: 10.1200/JOP.2016.018622

Chen, B., Liang, T., Yang, P., Wang, H., Liu, Y., Yang, F., et al. (2016). Classifying lower grade glioma cases according to whole genome gene expression. *Oncotarget* 7, 74031–74042. doi: 10.18632/oncotarget.12188

De Hoon, M. J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics* 20, 1453–1454. doi: 10.1093/bioinformatics/bth078

Dixit, K., and Raizer, J. (2017). Newer strategies for the management of low-grade gliomas. *Oncology (Williston Park, N. Y.)* 31, 680–682, 684–685.

Eckel-Passow, J. E., Lachance, D. H., Molinaro, A. M., Walsh, K. M., Decker, P. A., Sicotte, H., et al. (2015). Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N. Engl. J. Med.* 372, 2499–2508. doi: 10.1056/NEJMoa1407279

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, pl1. doi: 10.1126/scisignal.2004088

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Labak, C. M., Wang, P. Y., Arora, R., Guda, M. R., Asuthkar, S., Tsung, A. J., et al. (2016). Glucose transport: meeting the metabolic demands of cancer, and applications in glioblastoma treatment. *Am. J. Cancer Res.* 6, 1599–1608.

Liu, Z. G., Ding, J., Du, C., Xu, N., Wang, E. L., Li, J. Y., et al. (2018). Phosphoglycerate mutase 1 is highly expressed in C6 glioma cells and human astrocytoma. *Oncol. Lett.* 15, 8935–8940. doi: 10.3892/ol.2018.8477

Louis, D. N., Perry, A., Burger, P., Ellison, D. W., Reifenberger, G., Von Deimling, A., et al. (2014). International Society of Neuropathology—Haarlem consensus guidelines for nervous system tumor classification and grading. *Brain Pathol.* 24, 429–435. doi: 10.1111/bpa.12171

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Reitzer, L. J., Wice, B. M., and Kennell, D. (1979). Evidence that glutamine, not sugar, is the major energy source for cultured HeLa cells. *J. Biol. Chem.* 254, 2669–2676.

Saldanha, A. J. (2004). Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20, 3246–3248. doi: 10.1093/bioinformatics/bth349

Sanborn, J. Z., Salama, S. R., Grifford, M., Brennan, C. W., Mikkelsen, T., Jhanwar, S., et al. (2013). Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. *Cancer Res.* 73, 6036–6045. doi: 10.1158/0008-5472.CAN-13-0186

Smith, J. S., Perry, A., Borell, T. J., Lee, H. K., O'fallon, J., Hosek, S. M., et al. (2000). Alterations of chromosome arms 1p and 19q as predictors of survival in oligodendrogliomas, astrocytomas, and mixed oligoastrocytomas. *J. Clin. Oncol.* 18, 636–645. doi: 10.1200/JCO.2000.18.3.636

Soga, T. (2013). Cancer metabolism: key players in metabolic reprogramming. *Cancer Sci.* 104, 275–281. doi: 10.1111/cas.12085

The Cancer Genome Atlas Research Network, Mclendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., et al. (2008). Comprehensive genomic

characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061. doi: 10.1038/nature07385

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Warburg, O. (1956). On the Origin of Cancer Cells. *Science* 123, 309–314. doi: 10.1126/science.123.3191.309

Yan, H., Parsons, D. W., Jin, G., Mclendon, R., Rasheed, B. A., Yuan, W., et al. (2009). IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* 360, 765–773. doi: 10.1056/NEJMoa0808710

Zhang, M., Pan, Y., Qi, X., Liu, Y., Dong, R., Zheng, D., et al. (2018). Identification of new biomarkers associated with IDH mutation and prognosis in astrocytic tumors using NanoString nCounter Analysis System. *Appl. Immunohistochem. Mol. Morphol.* 26, 101–107. doi: 10.1097/PAI.0000000000000396

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 12 November 2019
doi: 10.3389/fonc.2019.01207

A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis

Guodong Yang[1†], *Yujiao Zhang*[2†] *and Jiyuan Yang*[1*]

[1] *Department of Oncology, The First People's Hospital Affiliated to Yangtze University, Jingzhou, China,* [2] *Respiratory Medicine, Huanggang Central Hospital Affiliated to Yangtze University, Huanggang, China*

Mounting evidence has demonstrated that a lot of miRNAs are overexpressed or downregulated in colorectal cancer (CRC) tissues and play a crucial role in tumorigenesis, invasion, and migration. The aim of our study was to screen new biomarkers related to CRC prognosis by bioinformatics analysis. By using the R language edgeR package for the differential analysis and standardization of miRNA expression profiles from The Cancer Genome Atlas (TCGA), 502 differentially expressed miRNAs (343 up-regulated, 159 down-regulated) were screened based on the cut-off criteria of $p < 0.05$ and $|\log2FC|>1$, then all the patients (421) with differentially expressed miRNAs and complete survival time, status were then randomly divided into train group (212) and the test group (209). Eight miRNAs with $p < 0.005$ were revealed in univariate cox regression analysis of train group, then stepwise multivariate cox regression was applied for constituting a five-miRNA (hsa-miR-5091, hsa-miR-10b-3p, hsa-miR-9-5p, hsa-miR-187-3p, hsa-miR-32-5p) signature prognostic biomarkers with obviously different overall survival. Test group and entire group shown the same results utilizing the same prescient miRNA signature. The area under curve (AUC) of receiver operating characteristic (ROC) curve for predicting 5 years survival in train group, test group, and whole cohort were 0.79, 0.679, and 0.744, respectively, which demonstrated better predictive power of prognostic model. Furthermore, Univariate cox regression and multivariate cox regression considering other clinical factors displayed that the five-miRNA signature could serve as an independent prognostic factor. In order to predict the potential biological functions of five-miRNA signature, target genes of these five miRNAs were analyzed by Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway and Gene Ontology (GO) enrichment analysis. The top 10 hub genes (ESR1, ADCY9, MEF2C, NRXN1, ADCY5, FGF2, KITLG, GATA1, GRIA1, KAT2B) of target genes in protein protein interaction (PPI) network were screened by string database and Cytoscape 3.6.1 (plug-in cytoHubba). In addition, 19 of target genes were associated with survival prognosis. Taken together, the current study showed the model of five-miRNA signature could efficiently function as a novel and independent prognosis biomarker and therapeutic target for CRC patients.

Keywords: microRNA, colorectal cancer, TCGA, prognosis, signature

OPEN ACCESS

Edited by:
Wan Zhu,
Stanford University, United States

Reviewed by:
Bryan R. G. Williams,
Hudson Institute of Medical Research, Australia
Ronald M. Przygodzki,
United States Department of Veterans Affairs, United States

****Correspondence:***
Jiyuan Yang
18163144297@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *30 July 2019*
Accepted: *23 October 2019*
Published: *12 November 2019*

Citation:
Yang G, Zhang Y and Yang J (2019) A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis. Front. Oncol. 9:1207. doi: 10.3389/fonc.2019.01207

INTRODUCTION

CRC is a very common gastrointestinal tumor with high incidence and mortality. It was estimated that more than 1.8 million new colorectal cancer cases and 0.88 million deaths will occur in 2018, accounting for about 1 in 10 cancer about incidence and mortality (1). CRC patients usually show a survival rate of <5 years due to early metastasis. Although treatments (such as surgery, radiotherapy, chemotherapy, and targeted therapy) have been developed fleetly, high recurrence, and poor prognosis remain troubling issues (2). Although various biomarkers have been discovered and were associated with the occurrence, progression and prognosis of colorectal cancer to date (3), their reliability remains controversial. Consequently, it is urgent to screen new potential diagnostic and prognostic biomarkers or therapeutic targets for CRC.

MicroRNAs (miRNAs), a vital component of the non-coding RNA family, are approximately made up of 18–25 nucleotides, which almost function via binding 3′ untranslated regions(UTR)or 5′UTR of mRNA to suppress translation and promote mRNA cleavage (4). Along with the advances of human genome-sequencing technology, a great number of miRNAs have been abundantly discovered. Increasing evidence demonstrated that miRNAs regulated various oncogenesis processes including cellular proliferation, angiogenesis, differentiation, and apoptosis by binding oncogenes or tumor suppresser genes (5). Zhang et al. displayed miRNA-519b-3p functioned as a tumor suppressor miRNA to suppress colorectal cancer cell proliferation and invasion by regulating the umtck/wnt signaling pathway (6). Wang et al. exhibited that miRNA-496 accelerated epithelial-mesenchymal transition and migration of CRC via targeting RASSF6, which was involved in Wnt-pathway (7). Huang et al. demonstrated miR-506 inhibited cell proliferation, invasion, and migration of CRC via reducing NR4A1 expression (8). Studies on miRNA in colorectal cancer are far more than that, there are also some studies on miRNA as prognostic factors, including single, and multiple combinations. Although TCGA database has been used to construct the miRNA signature prognostic models for colon cancer (9, 10), there are still some shortcomings with no miRNAs matures, model validation, and risk assessment.

In the present study, we constructed, verified and assessed a novel five-miRNA signature that predicted effectively over survival of CRC patients derived from TCGA database. Functional enrichment analysis revealed potential biological functions and signal pathways of five-miRNA signature associated with cancer, which enhances our understanding to molecular mechanisms of model in CRC.

MATERIALS AND METHODS

Data Download and Processing

The miRNA expression information [Case (455): Primary Site (Colon and Rectum), Program (TCGA),Project (TCGA-COAD and TCGA-READ), Disease Type (Adenomas and Adenocarcinomas); Files(473): Data Category (Transcriptome Profiling), Data Type (Isoform Expression Quantification)], mRNA expression information [Case (472): Primary Site (Colon and Rectum), Program (TCGA), Project (TCGA-COAD and TCGA-READ), Disease Type (Adenomas and Adenocarcinomas); Files (530): Data Category (Transcriptome Profiling), Data Type (Gene Expression Quantification)] and their related clinical information (476) (Data Category: Clinical, Data Format: BCR XML) (**Table 1**) of all colorectal cancer samples were downloaded from The Cancer Genome Atlas (TCGA) official website (https://cancergenome.nih.gov/) on July 3, 2019, the former of which contained 464 tumor samples and 9 normal samples, the latter included 488 tumor samples and 42 normal samples. The Fasta format sequences of all mature miRNA sequences (mature.fa) were downloaded from the miRBase website (http://www.mirbase.org/). We combined these two sets of data in the Perl language to obtain expression profile information for each mature miRNA.

TABLE 1 | Summary of patient cohort information.

Variables	Case	Percentage
GENDER		
Male	256	53.78%
Female	220	46.22%
AGE (YEARS)		
Range	31–90	
Median	68	14.29
RACE		
ASIAN	9	1.89%
BLACK	52	10.92%
WHITE	219	46.01%
Unknown	196	41.18%
CLINICAL STAGE		
Stage I	85	17.86%
Stage II	180	37.82%
Stage III	126	26.47%
Stage IV	70	14.71%
Unknown	15	3.15%
T STAGE		
T1+Tis	15	3.15%
T2	86	18.07%
T3	324	68.07%
T4	51	10.71%
LYMPH NODE STATUS		
N0	282	59.24%
N1	114	23.95%
N2	80	16.81%
Nx	1	0.21%
METASTATIC		
M0	355	74.58%
M1	69	14.50%
Mx	45	9.45%
Unknown	7	1.47%
CANCER TYPE		
COAD	385	80.88%
READ	91	19.12%

Identification of Differentially Expressed miRNAs, mRNA, and Their Combination With Patient Survival Data

We used R language 3.6.1 version edgeR package to compare the miRNA and mRNA expression of tumor group with normal group and normalize the expression profile of miRNAs and mRNA, whose mean value was >1, the screening criteria were corrected p value (FDR) <0.05 and |log2FC|>1 (11). We selected the clinical information of patients with survival time ≥30 days and combined it with differentially expressed and standardized miRNA and mRNA expression profiles.

Grouping of Samples and Construction, Validation, and Evaluation of Prognostic Models

We used the R language 3.6.1 version "caret" package to randomly divide the samples with complete survival information and differentially expressed miRNA expression profiles into two groups (train group and test group), and performed univariate Cox regression analysis of miRNAs for the train group.

In order to reduce the number of miRNAs with similar expression, miRNAs with p value< 0.005 were subjected to a stepwise multivariate Cox regression to construct the prognostic model. In the multivariate Cox regression analysis, we took advantage of the function of "Coxph" and "direction=both" in R language survival package (12). Then, the risk score of a prognostic miRNA signature comprising multiple miRNAs was established based on the summation of the product of each miRNA and its coefficient. Furthermore, we tested the Proportional Hazards Assumption in Cox model. This model was used to evaluate the survival prognosis of each patients in train group, test group, entire group using Kaplan-Meier curve, and log-rank test according to median value grouping of risk score, namely high risk group, and low risk group. The predictive power of the miRNA signature was assessed by calculating AUC of 3 years dependent ROC curve using "survivalROC" package (13).

Independent Prognostic Ability of the miRNA Signature Including Other Clinical Variables

The relationship between the prognostic miRNA signature and patients' overall survival was analyzed in the train

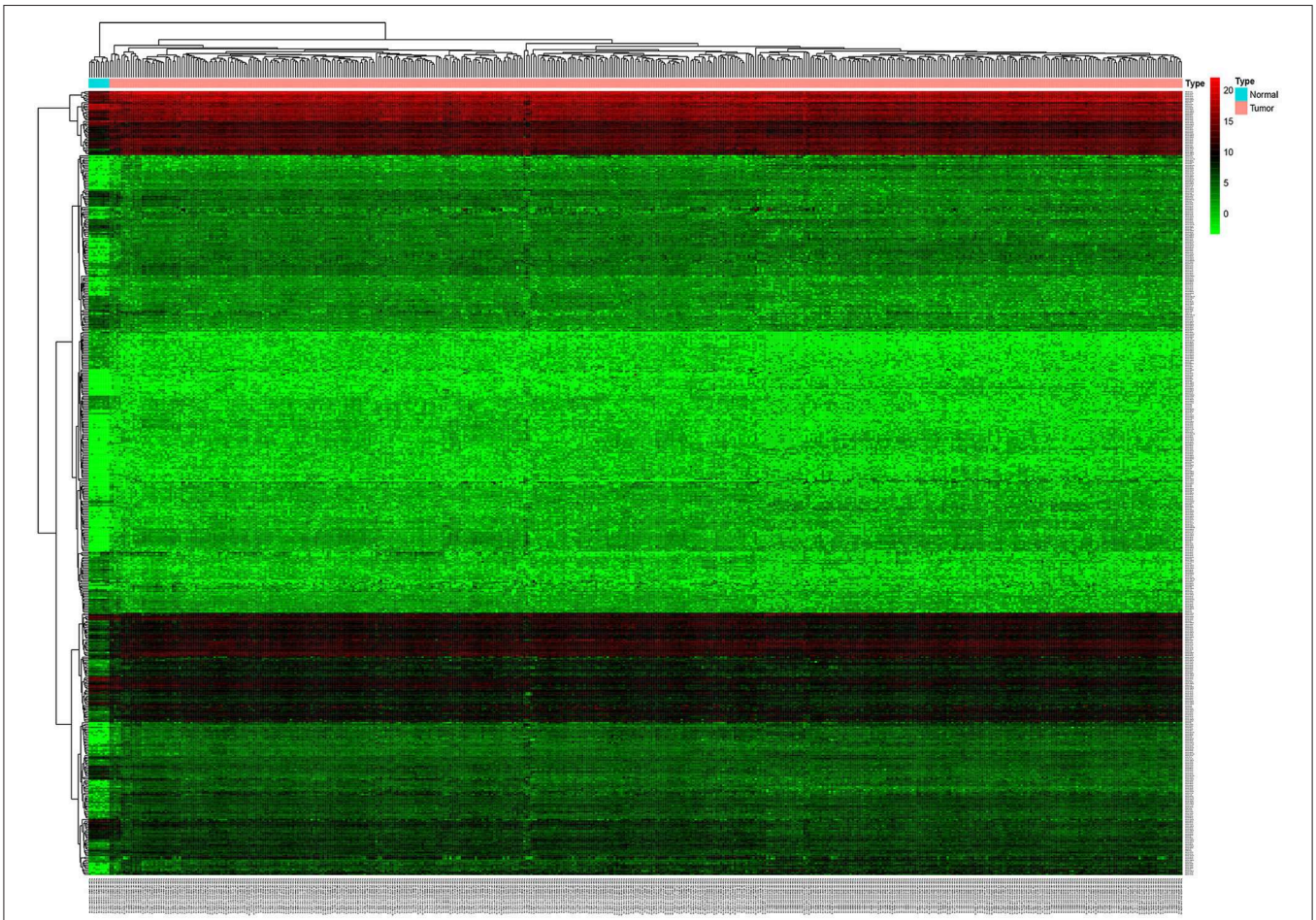


FIGURE 1 | Unsupervised hierarchical clustering heatmap based on the differentially expressed miRNAs between 464 colorectal cancer tissues and 9 normal tissues.

group by univariate Cox regression, as well as clinical variables (including age, gender, and clinical stage, lymph nodes, distant metastasis). Variables with p value < 0.05 in univariate Cox regression were further used for multivariate Cox regression analysis to determine whether they could function as independent prognostic factors. In order to compare the predictive power of this risk model compared to other clinical characteristics, we have drawn ROC curves for this model risk score and clinical characteristics. In addition, we tested the correlation of each miRNA to clinical features by using the SPSS 21.0 chi-square test, with a p-value of < 0.05 being considered meaningfully.

Target Genes Prediction of miRNA Signature and Their Potential Functions

We downloaded the miRNA prediction database from three miRNA target gene prediction websites including miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/), targetScan (http://www.targetscan.org) and miRDB (http://www.mirdb.org/), and used the Perl language to find the target genes of miRNA signature which are covered in at least 2 databases, meanwhile, utilizing the Venn diagram, and Cytoscape 3.6.1 to map the relationship between the miRNA and these target genes. To clarify whether the target genes of these miRNAs are likely to participate in the progression of colorectal cancer, we taken the intersection of these target genes and differentially expressed genes in colorectal cancer. All of these intersection genes obtained were analyzed by Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway and Gene Ontology (GO) enrichment analysis through the R language "clusterProfiler" package (14) and the "org.Hs.eg.db" package, The p adjust < 0.05 and q value < 0.05 was set as the cut-off criteria.

Screening of Hub Genes and Survival Related Gene

The PPI network of the STRING database (https://string-db.org/) (15) was applied to unearth the relationship between the target genes, the parameter of settings the medium confidence is 0.400. Then, the network relationship file was downloaded and the top 10 hub genes were identified in accordance with Cytoscape 3.6.1 and its plug-in (degrees ranking of cytoHubba). Meanwhile, The Kaplan-Meier method was used to check whether the intersection gene is related to over survival, log rank test < 0.05.

TABLE 2 | Univariate and multivariate Cox regression of differentially expressed miRNAs.

	Univariate Cox regression				Multivariate Cox regression				
id	HR	HR.95L	HR.95H	*P* value	Co ef	HR	HR.95L	HR.95H	*P* value
hsa-miR-485-5p	1.292	1.124	1.485	0.000					
hsa-miR-216a-5p	1.069	1.031	1.109	0.000					
hsa-miR-187-3p	1.044	1.019	1.069	0.000	0.031	1.031	1.001	1.062	0.041
hsa-miR-10b-3p	1.016	1.006	1.027	0.003	0.011	1.011	0.999	1.023	0.067
hsa-miR-32-5p	1.007	1.003	1.012	0.003	0.008	1.008	1.003	1.013	0.003
hsa-miR-9-5p	1.000	1.000	1.000	0.003	0.000	1.000	1.000	1.000	0.008
hsa-miR-5091	1.194	1.059	1.346	0.004	0.177	1.194	1.045	1.363	0.009
hsa-miR-5683	1.004	1.001	1.006	0.005					

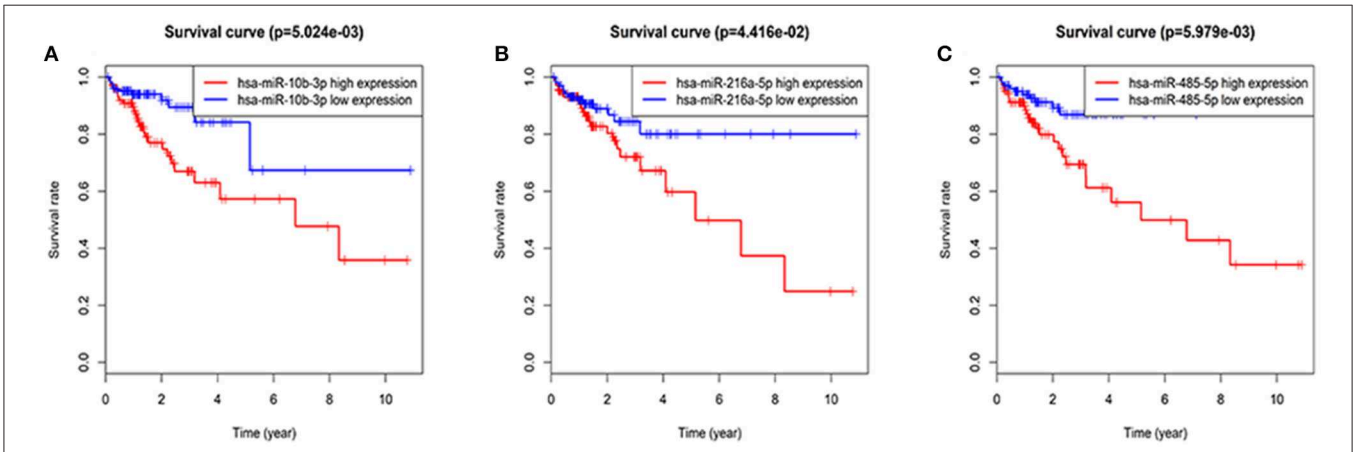


FIGURE 2 | Three miRNAs associated with overall survival in CRC patients using Kaplan–Meier curves and log-rank tests. The patients were stratified into high and low expression groups according to the median expression of each miRNA. **(A)** hsa-miR-10b-3p. **(B)** hsa-miR-216a-5p. **(C)** hsa-miR-485-5p.

Statistical Analysis

All statistical analyses are based on R language 3.6.1 version and attached packages.

RESULTS

Identification of Differentially Expressed miRNAs and mRNAs

Based on this screening criteria, miRNA mature expression profiles between 464 tumor samples and 9 normal samples showed 502 differentially expressed miRNAs (DEmiRNAs), of which 343 were up-regulated and 159 were down-regulated (**Figure 1**). mRNA expression profiles between 488 tumor samples and 42 normal samples showed 5,540 differentially expressed mRNAs (DEmRNAs), of which 2992 were up-regulated and 2,548 were down-regulated, displayed in **Supplemental Table 1**.

Construction of the Predictive Five-miRNA Signature

The entire group ($N = 421$) with miRNA mature expression profiles was randomly divided into train group ($N = 212$) (**Supplemental Table 2**) and test group ($N = 209$) (**Supplemental Table 3**). The univariate Cox regression analysis displayed that a total of thirty-two miRNAs were found to be associated with patients' overall survival (p value < 0.05) in the train group. For the reliability of the model, eight miRNAs (p value < 0.005) were selected for further analysis (**Table 2**). Kaplan-Meier method pointed out hsa-miR-10b-3p, hsa-miR-216a-5p, and hsa-miR-485-5p of eight miRNAs were associated with patients' overall survival (p value < 0.05; **Figure 2**), however,

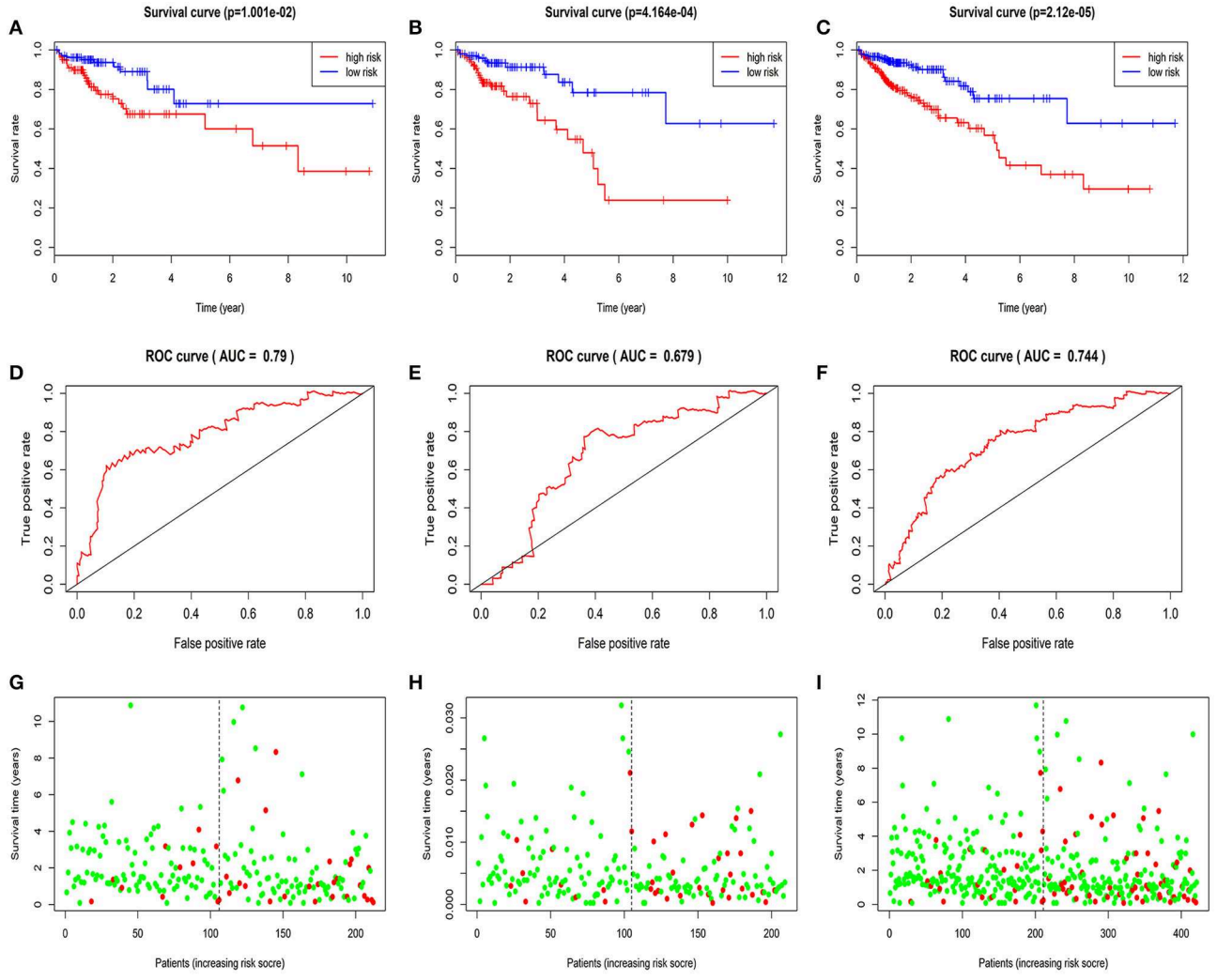


FIGURE 3 | Validation and evaluation of the predictive five-miRNA signature. Kaplan-Meier curves in the train group **(A)**, test group **(B)**, entire group **(C)**; The AUC of three years dependent curve in the train group **(D)**, test group **(E)**, entire group **(F)**, Survival status in high and low risk patients for train group **(G)**, test group **(H)**, entire group **(I)**, red dots represent death, green dots represent alive.

TABLE 3 | Univariate and multivariate Cox regression of clinical features.

Clinical features	Univariate Cox regression				Multivariate Cox regression			
	HR	HR.95L	HR.95H	*P* value	HR	HR.95L	HR.95H	*P* value
Age (continuous variable)	1.017	0.986	1.050	0.290				
Gender (male vs. female)	1.210	0.594	2.467	0.600				
Clinical stage (III+IV vs. I+II)	7.872	3.010	20.588	0.000	4.902	0.472	50.935	0.183
T stage (T3+4 vs. T1+2)	6.694	0.910	49.250	0.062				
M (M1 VS M0)	7.920	3.816	16.440	0.000	2.977	1.286	6.892	0.011
N (N1+2 vs. N0)	6.585	2.693	16.102	0.000	0.996	0.130	7.666	0.997
Five-miRNA signature	1.286	1.165	1.420	0.000	1.326	1.168	1.505	0.000

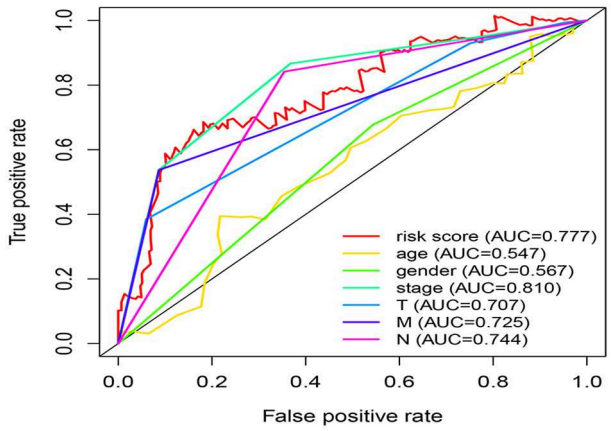


FIGURE 4 | Comparison of risk score and clinical features in predicting the accuracy of patients' survival prognosis.

the high expression of hsa-miR-485-5p with poor prognosis and the fact that hsa-miR-485-5p exhibited low expression in tumors is contradictory. Therefore, the remaining seven miRNAs were targeted for further analysis.

Based on the previous research, Five (hsa-miR-5091, hsa-miR-10b-3p, hsa-miR-9-5p, hsa-miR-187-3p, hsa-miR-32-5p) of the seven candidate miRNAs therein were finally screened out (**Table 2**) by stepwise multivariate Cox regression analysis, then a predictive miRNA signature model was established on the summation of the product of each miRNA and its coefficient in multivariate Cox regression as follows: miRNA signature risk score = (0.1769 × expression of hsa-miR-5091) + (0.0110 × expression of hsa-miR-10b-3p) + (0.0001 × expression of hsa-miR-9-5p) + (0.0305 × expression of hsa-miR-187-3p) + (0.0076 × expression of hsa-miR-32-5p). In addition, the results testing the Proportional Hazards Assumption in Cox model demonstrated that all the *P* values are higher than 0.05, which means that they meet the PH test (**Supplemental Table 4**).

Prediction of the Five-miRNA Signature for Over Survival in the Train Group, Test Group, and Entire Group

Based on median value grouping of risk score. Kaplan-Meier curves shown high risk group had an obviously poorer overall survival compared to low risk group in the train group (p = 1.001E-02), test group (p = 4.164E-04) and entire group (p = 2.12E-05; **Figures 3A–C**). The train group shown overall survival of 5 years for patients with high and low risk group were 60.0 and 72.8%, respectively. The test group demonstrated that overall survival of 5 years for patients with high and low risk group were 39.9 and 62.7%, respectively. The entire group displayed that overall survival of 5 years for patients with high and low risk group were 53.0 and 62.8%, respectively.

Evaluation of the Five-miRNA Signature for Over Survival in the Train Group, Test Group, and Entire Group

The AUC of 3 years dependent ROC for the five-miRNA signature achieved 0.790, 0.679, 0.744, respectively, in the train group, test group and entire group (**Figures 3D–F**), which demonstrated the better performance of model in predicting CRC patient survival risk. In addition, in the three groups, the patients with high risk score had higher mortality rates than low (**Figures 3G–I**).

Independence of the Five-miRNA Signature Considering Other Clinical Factors

Univariate Cox regression analysis exhibited that the five-miRNA signature was evidently associated with patients' overall survival (hazard ratio HR = 1.286, confidence interval 95% CI = 1.164–1.420, p = 6.719E-07; **Table 3**). Multivariate Cox regression analysis pointed out that the five-miRNA signature remained independent with overall survival considering other conventional clinical factors (HR = 1.326, 95% CI = 1.168–1.505, p = 1.23E-05), such as clinical stage, T stage, Lymph-node status, distant metastasis, which makes it possible to be a prognostic marker for CRC in the future. Meanwhile, distant metastasis was also found to be an independent prognostic factor (HR = 2.976, 95% CI = 1.285–6.891, p = 0.01). The ROC curves for this model risk score and clinical characteristics demonstrated that risk score (0.777), clinical stage (0.810), T stage (0.707), Lymph-node status (0.725), and distant metastasis (0.744) had a high predictive ability (**Figure 4**). In addition, the results about the correlation of each miRNA to clinical features demonstrated hsa-miR-10b-3p was associated with T stage (p = 0.011), hsa-miR-9-5p was associated with age (p = 0.032), and clinical stage (p = 0.049), hsa-mir-3189 was associated with Metastasis (p = 0.002) and clinical stage (p = 0.042; **Table 4**), which further

TABLE 4 | The correlation of each miRNA to clinical features.

Variables	Numbers	hsa-miR-5091		χ^2 test *P* value	hsa-miR-10b-3p		χ^2 test *P* value	hsa-miR-9-5p		χ^2 test P value	hsa-miR-187-3p		χ^2 test *P* value	hsa-mir-3189		χ^2 test *P* value
		Low Expression	High Expression		Low Expression	High Expression		Low Expression	High Expression		Low Expression	High Expression		Low Expression	High Expression	
GENDER																
Female	81	46	35	0.12	43	38	0.425	44	37	0.334	40	41	0.834	41	40	0.4
Male	108	49	59		51	57		51	57		55	53		48	60	
AGE AT DIAGNOSIS																
>60	136	69	67	0.836	66	70	0.595	75	61	0.032	72	64	0.238	59	77	0.102
≤60	53	26	27		28	25		20	33		23	30		30	23	
T STAGE																
T1+2	40	23	17	0.303	27	13	0.011	21	19	0.75	25	15	0.081	20	20	0.678
T3+4	149	72	77		67	82		74	75		70	79		69	80	
METASTASIS																
M0	155	78	77	0.973	79	76	0.469	81	74	0.242	83	72	0.054	81	74	0.002
M1	34	17	17		15	19		14	20		12	22		8	26	
LYMPH NODE STATUS																
N0	103	53	60	0.72	54	49	0.418	58	45	0.069	55	48	0.346	54	49	0.108
N1-2	86	42	44		40	46		37	49		40	46		35	51	
STAGE																
I+II	100	50	50	0.939	52	48	0.509	57	43	0.049	52	48	0.613	54	46	0.044
III+IV	89	45	44		42	47		38	51		43	46		35	54	

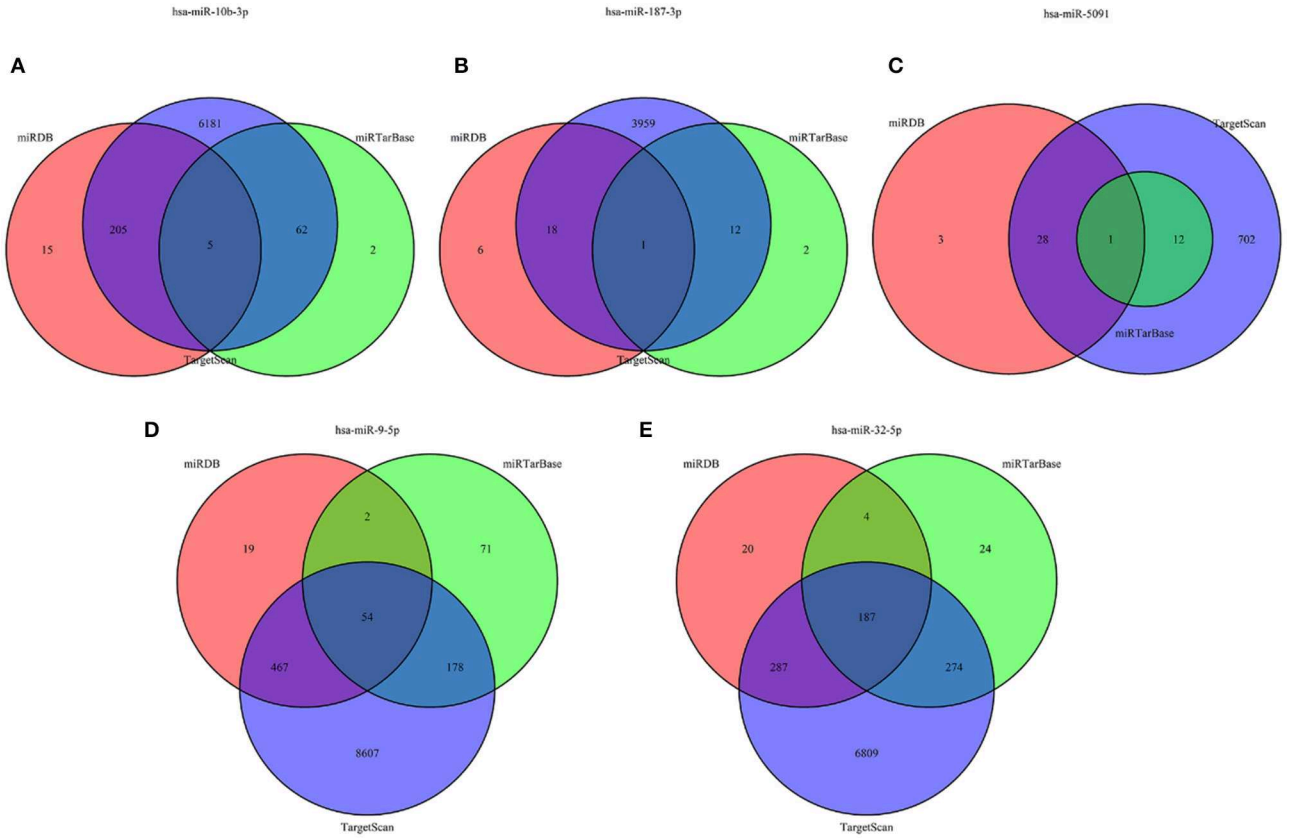


FIGURE 5 | Venn diagram of target genes for five miRNAs. **(A)** hsa-miR-10b-3p, **(B)** hsa-miR-187-3p, **(C)** hsa-miR-5091, **(D)** hsa-miR-9-5p, **(E)** hsa-miR-32-5p.

suggested that these miRNAs do have a close relationship with some clinical features.

Prediction of Target Genes for the Five miRNAs

The target genes regulated by the five miRNAs, were predicted in at least 2 databases. To further enhance the reliability of the bioinformatic analysis, the overlapping target genes were identified. The results indicated that 41, 272, 701, 31, and 752 overlapping genes were identified for hsa-miR-5091, hsa-miR-10b-3p, hsa-miR-9-5p, hsa-miR-187-3p, hsa-miR-32-5p, respectively, by the three databases above, which were shown using Venn diagram (**Figure 5**) and network map of miRNA-target genes (**Supplemental Figure 1**). A total of 1,672 target genes was predicted for the five miRNAs. To clarify whether the target genes of these miRNAs are likely to participate in the progression of CRC, the above obtained 5540 DEmRNAs (up-regulated 2992, down-regulated 2548) was used for analysis. The intersection of target mRNAs for down-regulated miRNAs (hsa-miR-5091, hsa-miR-187-3p) and upregulated mRNAs, and target mRNAs for upregulated miRNAs (hsa-miR-32-5p, hsa-miR-10b-3p, hsa-miR-9-5p) and downregulated mRNAs were taken. The results were performed on a total of 246 genes including 12 up-regulated genes, 234 down-regulated genes, respectively (**Supplemental Figure 2**). The sub network between the five miRNAs and their 246 target genes was shown in **Figure 6**.

Functional Enrichment Analysis of Target Genes Associated CRC

The results of GO annotation about the target genes associated CRC are 234 (**Supplemental Table 5**). The top fifteen terms from the GO results: biological process (BP), cellular component (CC), and molecular function (MF) were demonstrated in dotplot (**Figures 7A–C**). In the three categories, BP analysis mostly include axon development, axonogenesis, and stem cell differentiation, CC analysis was mainly contained synaptic membrane, postsynaptic membrane and neuronal cell body, MF analysis mainly contained metal ion transmembrane transporter activity, transcriptional activator activity and DNA binding, ion channel binding. The results of KEGG pathways about the target genes associated CRC are 18 (**Table 5**), of which counts $>$ 10 were mainly enriched in the cGMP-PKG signaling pathway, cAMP signaling pathway, Calcium signaling pathway, Neuroactive ligand-receptor interaction In addition, to provide a readable graphic representation of the complex relationship between target genes and relative KEGG pathway, the "pathway-gene

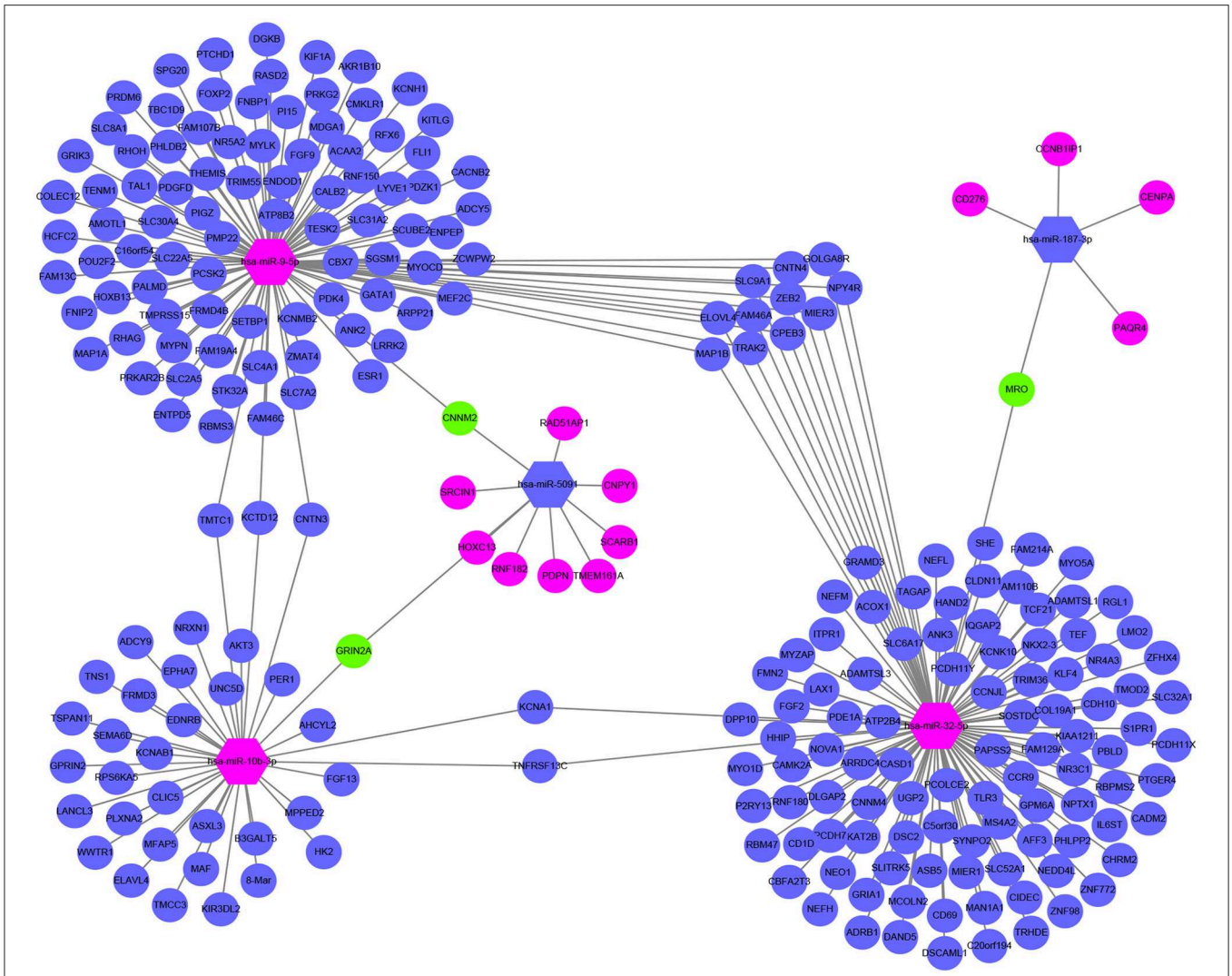


FIGURE 6 | Sub network map of miRNA regulating mRNA. The hexagon represents miRNA. The circle stands for mRNA. Red means upregulated, blue means downregulated, and green means both.

network" and "pathway-pathway network" was also shown in **Figures 7D–F**.

Hub Genes of PPI Network and Survival Related Target Genes

Total of 244 of the 246 target genes were filtered into the target genes PPI network complex, containing 178 nodes and 326 edges, 10 hub gene (ESR1, ADCY9, MEF2C, NRXN1, ADCY5, FGF2, KITLG, GATA1, GRIA1, KAT2B) were screened according to Cytoscape 3.6.1 and its plug-in (degree ranking of cytoHubba) (**Figure 8** and **Table 6**). In addition, Kaplan-Meier method showed that the expression of 18 of the 246 genes (AHCYL2, AKR1B10, CBFA2T3, CCNJL, CCR9, CLIC5, DPP10, FAM46C, GATA1, IQGAP2, MAN1A1, MIER1, NR5A2, PHLPP2, PTGER4, RBM47, RPS6KA5, TSPAN11) were positively associated with survival prognosis, however, the high expression of SRCIN1 shown a poorer over survival (**Figure 9**).

DISCUSSION

Colorectal cancer is a highly malignant tumor, which is particularly prone to liver and lung metastasis, seriously affecting the survival prognosis of patients (16). Therefore, finding a prognostic marker with high specificity and sensitivity is becoming more and more urgent for patients. Extensive evidence displayed miRNAs can regulate the expression of abundant genes, playing critical roles in many biological processes of human malignant tumor (17). Especially, recent studies have revealed that distinct miRNA-expression profiles seriously affected the development and progression of CRC (18, 19). At present, several miRNAs are known to be used as potential prognostic

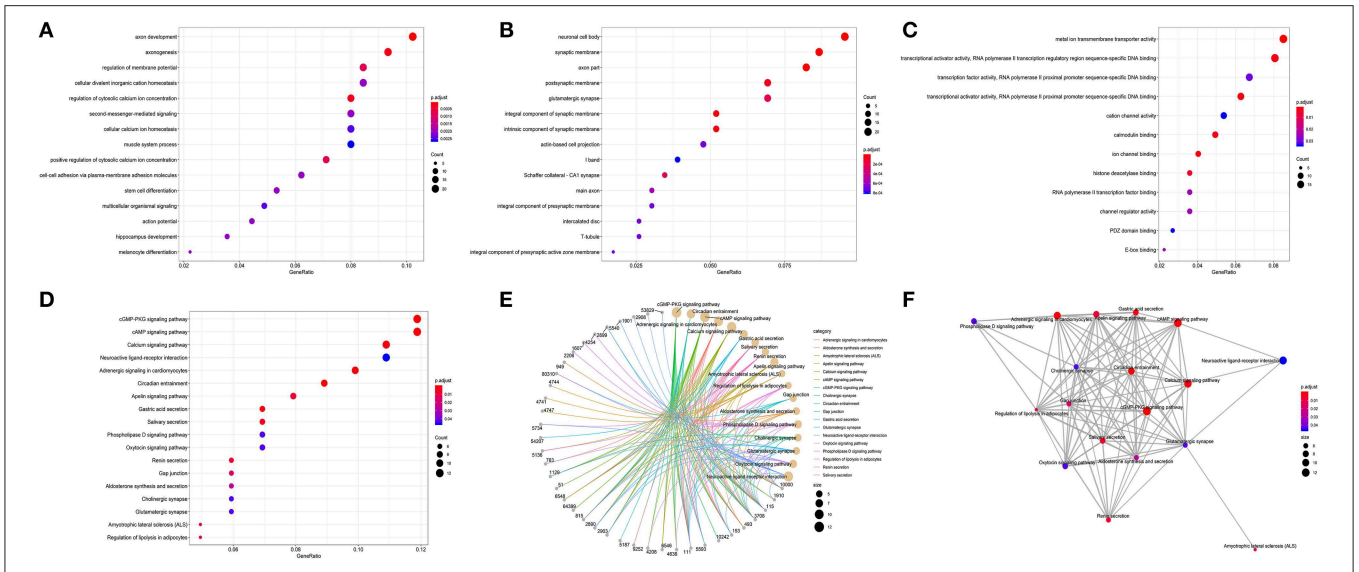


FIGURE 7 | Functional enrichment analysis of target genes associated CRC. **(A)** BP, **(B)** CC, **(C)** MF, **(D)** dotplot of KEGG signal pathway shown the counts of genes, **(E)** cnetplot of KEGG signal pathway shown the "pathway-gene" network, **(F)** emapplot of KEGG signal pathway shown the "pathway-pathway" network.

TABLE 5 | KEGG pathways of target genes associated CRC.

ID	Description	P Adjust	Q value	Count	Gene ID
hsa04022	cGMP-PKG signaling pathway	0.00028	0.000212	12	AKT3/EDNRB/ADCY9/ITPR1/ATP2B4/ADRB1/KCNMB2/PRKG2/ADCY5/MYLK/SLC8A1/MEF2C
hsa04713	Circadian entrainment	0.000397	0.0003	9	RPS6KA5/PER1/GRIN2A/ADCY9/ITPR1/GRIA1/CAMK2A/PRKG2/ADCY5
hsa04024	cAMP signaling pathway	0.001032	0.000782	12	AKT3/GRIN2A/ADCY9/ATP2B4/HHIP/SLC9A1/GRIA1/CAMK2A/ACOX1/ADRB1/CHRM2/ADCY5
hsa04261	Adrenergic signaling in cardiomyocytes	0.001032	0.000782	10	RPS6KA5/AKT3/ADCY9/ATP2B4/SLC9A1/CAMK2A/ADRB1/ADCY5/SLC8A1/CACNB2
hsa04020	Calcium signaling pathway	0.001443	0.001093	11	GRIN2A/EDNRB/ADCY9/ITPR1/ATP2B4/CAMK2A/ADRB1/PDE1A/CHRM2/MYLK/SLC8A1
hsa04971	Gastric acid secretion	0.001615	0.001223	7	ADCY9/ITPR1/SLC9A1/KCNK10/CAMK2A/ADCY5/MYLK
hsa04970	Salivary secretion	0.004453	0.003373	7	ADCY9/ITPR1/ATP2B4/SLC9A1/ADRB1/PRKG2/ADCY5
hsa04924	Renin secretion	0.006429	0.00487	6	PTGER4/ITPR1/ADRB1/PDE1A/PRKG2/ADCY5
hsa04371	Apelin signaling pathway	0.008483	0.006426	8	AKT3/ADCY9/ITPR1/SLC9A1/ADCY5/MYLK/SLC8A1/MEF2C
hsa05014	Amyotrophic lateral sclerosis (ALS)	0.009933	0.007525	5	GRIN2A/NEFL/NEFM/NEFH/GRIA1
hsa04923	Regulation of lipolysis in adipocytes	0.012842	0.009728	5	AKT3/ADCY9/ADRB1/PRKG2/ADCY5
hsa04540	Gap junction	0.015926	0.012064	6	ADCY9/ITPR1/ADRB1/PDGFD/PRKG2/ADCY5
hsa04925	Aldosterone synthesis and secretion	0.025762	0.019515	6	ADCY9/ITPR1/ATP2B4/CAMK2A/ADCY5/SCARB1
hsa04072	Phospholipase D signaling pathway	0.04312	0.032663	7	AKT3/ADCY9/MS4A2/DGKB/PDGFD/ADCY5/KITLG
hsa04725	Cholinergic synapse	0.04312	0.032663	6	AKT3/ADCY9/ITPR1/CAMK2A/CHRM2/ADCY5
hsa04724	Glutamatergic synapse	0.04312	0.032663	6	GRIN2A/ADCY9/ITPR1/GRIA1/ADCY5/GRIK3
hsa04921	Oxytocin signaling pathway	0.04312	0.032663	7	ADCY9/ITPR1/CAMK2A/ADCY5/MYLK/MEF2C/CACNB2
hsa04080	Neuroactive ligand-receptor interaction	0.046345	0.035106	11	GRIN2A/EDNRB/PTGER4/NPY4R/GRIA1/S1PR1/NR3C1/P2RY13/ADRB1/CHRM2/GRIK3

indictors in various cancers, including miR-191 (20), miR-1908 (21), miR-200c (22), and miR-217 (23). However, overwhelming studies manifested that multiple miRNA signature have bigger advantages than single miRNA on the hand of statistically robust analysis. Thence before our study, there have been a lot of prognostic markers based on multiple miRNA signature in tumors (24–26), especially colorectal cancer (9, 10, 27). There are many differences between our research and previous studies yet, such as research methods, sample size, and most importantly, we use miRNA matures and sample groupings to validate the model.

In the current study, we download mature miRNA expression profiles and corresponding patients' clinical information of CRC

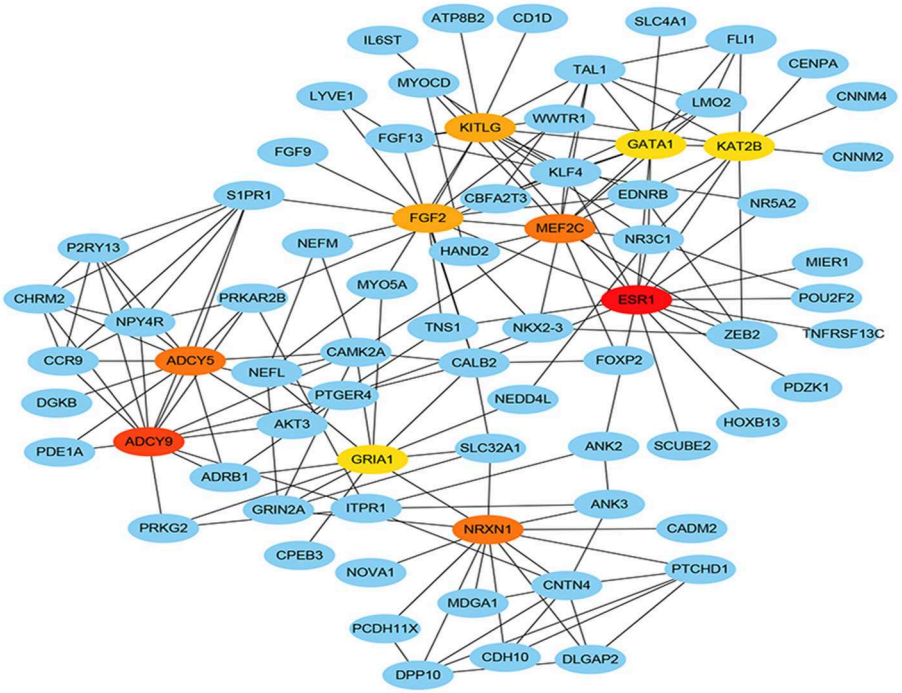


FIGURE 8 | Hub genes of PPI network. The darker the color, the bigger the degrees.

TABLE 6 | Identification of hub genes by cytoHubba.

Node_name	MCC	DMNC	MNC	Degree	EPC	Bottle Neck	Ec Centricity	Closeness	Radiality	Betweenness	Stress	Clustering Coefficient
ESR1	40	0.238	11	17	57.77	46	0.101	67.513	10.95	7030.605	17450	0.103
ADCY9	742	0.282	14	14	56.906	4	0.113	59.613	10.577	1678.562	6756	0.275
MEF2C	38	0.255	11	13	57.288	14	0.113	61.663	10.735	2113.807	7622	0.192
NRXN1	44	0.321	8	13	44.099	20	0.113	54.98	10.317	3224.563	11316	0.179
ADCY5	739	0.337	12	13	56.085	13	0.113	56.846	10.453	1075.731	4490	0.295
FGF2	19	0.256	7	12	57.027	15	0.101	63.513	10.826	2905.838	9030	0.106
KITLG	32	0.321	8	12	56.272	9	0.09	60.182	10.639	1839.913	5646	0.167
GATA1	67	0.419	10	11	56.982	8	0.101	58.69	10.566	1203.977	3882	0.382
GRIA1	22	0.329	7	11	54.01	26	0.113	61.78	10.803	3086.888	10038	0.164
KAT2B	32	0.402	7	11	53.712	9	0.101	57.856	10.498	1381.704	4194	0.2

from TCGA database. By using the R language edgeR package for the differential analysis, 502 DEmiRNAs were obtained. All the patients were randomly divided into train group and test group, then a five-miRNA signature model (hsa-miR-5091, hsa-miR-10b-3p, hsa-miR-9-5p, hsa-miR-187-3p, hsa-miR-32-5p) was constructed by univariate Cox regression and stepwise multivariate Cox regression in train group. Meanwhile, a five-miRNA signature was validated in test group and entire group. Based on median value grouping of risk score. Kaplan-Meier curves shown high risk group had an obviously poorer overall survival compared to low risk group in the three group. Evaluation of the five-miRNA signature for over survival in the three group by ROC curve displayed better predictive power. Univariate Cox regression and multivariate Cox regression analysis also pointed out that the five-miRNA signature remained independent with overall survival considering other conventional clinical factors for CRC patients. Most of these five miRNAs have been reported to participate in the research progress of various tumors. Lu et al. demonstrated that the expression level of mir-10b-3p was obviously upregulated in tumor and serum samples of esophageal cancer (ESCC) patients. The expression level of mir-10b-3p is not only correlated with lymph node metastasis and clinical staging, but also serves as an independent prognostic biomarker for overall survival of ESCC patients. Augmented expression of mir-10b-3p stimulates cell proliferation, invasion, and migration through directly combining the FOXO3 3'UTR

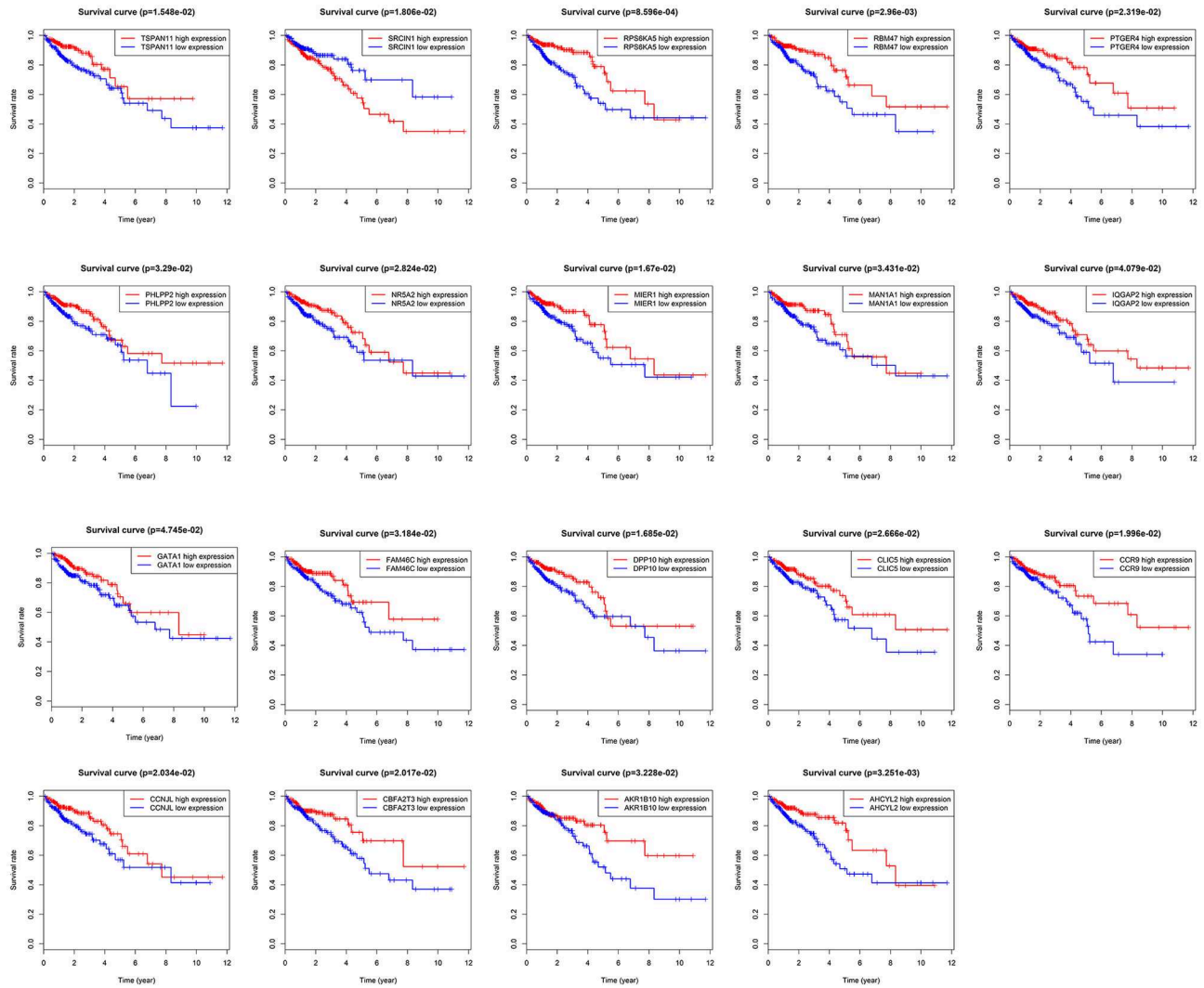


FIGURE 9 | Target genes associated with over survival.

in ESCC (28). Chen et al. shown that miR-9-5p expression was upregulated in prostate cancer cells, functioned as oncogene role in the proliferation, migration, invasion, and epithelial-mesenchymal transition (EMT) of prostate cancer cells by binding StarD13 (29). Dou et al. demonstrated that miR-187-3p was lowly expressed in hepatic carcinoma (HCC) tissues and cell lines, and was not only correlated with clinical stage and metastasis of HCC, but also accelerated effects of hypoxia on EMT of HCC cells. Furthermore, miR-187-3p suppressed EMT process in HCC via regulating S100A4 (30). Fu et al. reported that miR-32-5p was markedly upregulated in the HCC multidrug-resistant cell line (Bel/5-FU). Overexpression of miR-32-5p demonstrated a worse prognosis, miR-32-5p regulated the PI3K/Akt pathway via inhabiting PTEN and leaded to multidrug resistance by exosomes, then advanced epithelial-mesenchymal transition (EMT) and angiogenesis (31). However, the current research mechanism of hsa-miR-5091 in tumors has not been reported yet, so more experiments in the future need to be carried out to hsa-miR-5091, especially in CRC.

To further understand the regulatory mechanism of the five-miRNA signature in colorectal cancer, the target genes of five miRNAs in the model were predicted by three target gene prediction databases. At the same time, based on the study of colorectal cancer, we obtained the intersection of the target genes of these miRNAs and the differentially expressed genes from the TCGA database, and performed functional enrichment analysis on these intersection genes. The GO annotation of the target genes was mainly associated with axon development, axonogenesis and stem cell differentiation, synaptic membrane, postsynaptic membrane, and neuronal cell body, metal ion transmembrane transporter activity, transcriptional activator activity and DNA binding, ion channel binding. The signal pathways of the target genes mainly enriched in the cGMP-PKG signaling pathway, cAMP signaling pathway, Calcium signaling pathway, Neuroactive ligand-receptor interaction. Ren et al. illuminated that the cGMP/PKG signaling pathway played an essential role on proliferation and survival of human renal carcinoma cells (32). Park et al. displayed that the cAMP

signaling pathway regulated by the Epac-Rap1-Akt pathway caused suppression of JNK-dependent HDAC8 degradation, which augments cisplatin-induced apoptosis by inhabiting TIPRL expression in lung cancer cells (33). Monteith GR reviewed that calcium signaling pathway not only played key role on proliferation, invasion and sensitivity to cell death, but also in the establishment and maintenance of multidrug resistance and the tumor microenvironment (34). These signaling pathways show their effects on tumors to varying degrees, and these three signaling pathways are only the tip of the iceberg of the target gene involved in signaling pathway, which prompts that our constructed miRNA prognosis model may be involved in the regulation of tumor signaling pathways.

In order to find key nodes of the miRNA signature model regulating colorectal cancer 10 hub genes (ESR1, ADCY9, MEF2C, NRXN1, ADCY5, FGF2, KITLG, GATA1, GRIA1, KAT2B) were screened according to Cytoscape 3.6.1 and its plug-in (degree ranking of cytoHubba). In addition, the Kaplan-Meier method showed that the expression of 18 genes (AHCYL2, AKR1B10, CBFA2T3, CCNJL, CCR9, CLIC5, DPP10, FAM46C, GATA1, IQGAP2, MAN1A1, MIER1, NR5A2, PHLPP2, PTGER4, RBM47, RPS6KA5, TSPAN11) were positively associated with survival prognosis, however the high expression of SRCIN1 shown a poorer over survival. Surprisingly, GATA1 (GATA binding protein 1) is not only a key gene in the PPI network, but also related to over survival of patients, which encodes s a protein which belongs to the GATA family of transcription factors and promoted erythroid development via adjusting the switch of fetal hemoglobin to adult hemoglobin. Wang et al. pointed out that decreased of GATA-1 was to the benefit of high expression of IRF-3 in lung adenocarcinoma cells by binding with a specific domain of IRF-3 promoter, consequently, alternating the immunomodulatory function in tumorigenesis (35). Thus, the miRNA signature may affect the survival prognosis of colorectal cancer patients and the colorectal cancer progression through regulating GATA1.

CONCLUSION

In summary, our study not only constructed a new predictive model of miRNA signature prognosis through miRNA mature expression profiling, but also by grouping to verify and evaluating the predictive ability of the model, the most important thing is that it can be used as an independent prognostic factors in CRC. In addition, the potential function is inferred by predicting the target genes of the model, which enhance our comprehension to tumorigenesis and progression of CRC. However, this is just a study based on the TCGA database using bioinformatics. We hope that there will be other databases and a large number of experiments to verify the feasibility of this prognostic model in the future and provide a reliable predictor and therapeutic target for CRC patients.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

YZ downloaded the miRNA and mRNA expression information. GY constructed miRNA signature model and performed the statistical analysis using R language software, and wrote the first draft of the manuscript. JY contributed conception and design of the study and checked the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.01207/full#supplementary-material

Supplemental Figure 1 | The network map between miRNAs and target genes. The hexagon represents miRNA, the circle stands for mRNA. Red means upregulated, blue means downregulated.

Supplemental Figure 2 | The intersection of target mRNAs for miRNA and differentially expressed mRNAs.

Supplemental Table 1 | Differentially expressed mRNAs between colorectal cancer samples and normal samples.

Supplemental Table 2 | GO annotation of the target genes.

Supplemental Table 3 | Survival information of differentially expressed miRNA train group.

Supplemental Table 4 | Proportional Hazards Assumption in Cox model.

Supplemental Table 5 | GO annotation of the target genes.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2018) 68:394–424. doi: 10.3322/caac.21492
2. Al Bandar MH, Kim NK. Current status and future perspectives on treatment of liver metastasis in colorectal cancer. *Oncol Rep*. (2017) 37:2553–64. doi: 10.3892/or.2017.5531
3. Gires O. Lessons from common markers of tumor-initiating cells in solid cancers. *Cell Mol Life Sci*. (2011) 68:4009–22. doi: 10.1007/s00018-011-0772-9
4. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. *Nature*. (2004) 431:343–9. doi: 10.1038/nature02873
5. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. (2004) 116:281–97. doi: 10.1016/S0092-8674(04)00045-5
6. Zhang Y, Sun M, Chen Y, Li B. MiR-519b-3p inhibits the proliferation and invasion in colorectal cancer via modulating the uMtCK/Wnt signaling pathway. *Front Pharmacol*. (2019) 10:741. doi: 10.3389/fphar.2019.00741
7. Wang H, Yan B, Zhang P, Liu S, Li Q, Yang J, et al. MiR-496 promotes migration and epithelial-mesenchymal transition by targeting RASSF6 in colorectal cancer. *J Cell Physiol*. (2019). doi: 10.1002/jcp.29066. [Epub ahead of print].
8. Huang M, Xie X, Song X, Gu S, Chang X, Su T, et al. MiR-506 suppresses colorectal cancer development by inhibiting orphan nuclear receptor NR4A1 expression. *J Cancer*. (2019) 10:3560–70. doi: 10.7150/jca.28272

9. Xu J, Zhao J, Zhang R. Four microRNAs signature for survival prognosis in colon cancer using TCGA data. *Sci Rep*. (2016) 6:38306. doi: 10.1038/srep38306
10. Wei HT, Guo EN, Liao XW, Chen LS, Wang JL, Ni M, et al. Genomescale analysis to identify potential prognostic microRNA biomarkers for predicting overall survival in patients with colon adenocarcinoma. *Oncol Rep*. (2018) 40:1947–58. doi: 10.3892/or.2018.6607
11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
12. Stel VS, Dekker FW, Tripepi G, Zoccali C, Jager KJ. Survival analysis II: cox regression. *Nephron Clin Pract*. (2011) 119:c255–60. doi: 10.1159/000328916
13. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. (2000) 56:337–44. doi: 10.1111/j.0006-341X.2000.00337.x
14. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. (2012) 16:284–7. doi: 10.1089/omi.2011.0118
15. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. (2017) 45:D362–D8. doi: 10.1093/nar/gkw937
16. Adam R, de Gramont A, Figueras J, Kokudo N, Kunstlinger F, Loyer E, et al. Managing synchronous liver metastases from colorectal cancer: a multidisciplinary international consensus. *Cancer Treat Rev*. (2015) 41:729–41. doi: 10.1016/j.ctrv.2015.06.006
17. Bertoli G, Cava C, Castiglioni I. MicroRNAs: new biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics*. (2015) 5:1122–43. doi: 10.7150/thno.11543
18. Ding L, Lan Z, Xiong X, Ao H, Feng Y, Gu H, et al. The dual role of MicroRNAs in colorectal cancer progression. *Int J Mol Sci*. (2018) 19:2791. doi: 10.3390/ijms19092791
19. Masuda T, Hayashi N, Kuroda Y, Ito S, Eguchi H, Mimori K. MicroRNAs as biomarkers in colorectal cancer. *Cancers*. (2017) 9:E124. doi: 10.3390/cancers9090124
20. Gao X, Xie Z, Wang Z, Cheng K, Liang K, Song Z. Overexpression of miR-191 predicts poor prognosis and promotes proliferation and invasion in esophageal squamous cell carcinoma. *Yonsei Med J*. (2017) 58:1101–10. doi: 10.3349/ymj.2017.58.6.1101
21. Teng C, Zheng H. Low expression of microRNA-1908 predicts a poor prognosis for patients with ovarian cancer. *Oncol Lett*. (2017) 14:4277–81. doi: 10.3892/ol.2017.6714
22. Si L, Tian H, Yue W, Li L, Li S, Gao C, et al. Potential use of microRNA-200c as a prognostic marker in non-small cell lung cancer. *Oncol Lett*. (2017) 14:4325–30. doi: 10.3892/ol.2017.6667
23. Yang J, Zhang HF, Qin CF. MicroRNA-217 functions as a prognosis predictor and inhibits pancreatic cancer cell proliferation and invasion via targeting E2F3. *Eur Rev Med Pharmacol Sci*. (2017) 21:4050–7.
24. Liang B, Zhao J, Wang X. A three-microRNA signature as a diagnostic and prognostic marker in clear cell renal cancer: an *in silico* analysis. *PLoS ONE*. (2017) 12:e0180660. doi: 10.1371/journal.pone.0180660
25. Zhang C, Zhang CD, Ma MH, Dai DQ. Three-microRNA signature identified by bioinformatics analysis predicts prognosis of gastric cancer patients. *World J Gastroenterol*. (2018) 24:1206–15. doi: 10.3748/wjg.v24.i11.1206
26. Shi XH, Li X, Zhang H, He RZ, Zhao Y, Zhou M, et al. A five-microRNA signature for survival prognosis in pancreatic adenocarcinoma based on TCGA data. *Sci Rep*. (2018) 8:7638. doi: 10.1038/s41598-018-22493-5
27. Zanutto S, Ciniselli CM, Belfiore A, Lecchi M, Masci E, Delconte G, et al. Plasma miRNA-based signatures in CRC screening programs. *Int J Cancer*. (2019). doi: 10.1002/ijc.32573. [Epub ahead of print].
28. Lu YF, Yu JR, Yang Z, Zhu GX, Gao P, Wang H, et al. Promoter hypomethylation mediated upregulation of MicroRNA-10b-3p targets FOXO3 to promote the progression of esophageal squamous cell carcinoma. (ESCC). *J Exp Clin Cancer Res*. (2018) 37:301. doi: 10.1186/s13046-018-0966-1
29. Chen L, Hu W, Li G, Guo Y, Wan Z, Yu J. Inhibition of miR-9–5p suppresses prostate cancer progress by targeting StarD13. *Cell Mol Biol Lett*. (2019) 24:20. doi: 10.1186/s11658-019-0145-1
30. Dou C, Liu Z, Xu M, Jia Y, Wang Y, Li Q, et al. miR-187–3p inhibits the metastasis and epithelial-mesenchymal transition of hepatocellular carcinoma by targeting S100A4. *Cancer Lett*. (2016) 381:380–90. doi: 10.1016/j.canlet.2016.08.011
31. Fu X, Liu M, Qu S, Ma J, Zhang Y, Shi T, et al. Exosomal microRNA-32–5p induces multidrug resistance in hepatocellular carcinoma via the PI3K/Akt pathway. *J Exp Clin Cancer Res*. (2018) 37:52. doi: 10.1186/s13046-018-0677-7
32. Ren Y, Zheng J, Yao X, Weng G, Wu L. Essential role of the cGMP/PKG signaling pathway in regulating the proliferation and survival of human renal carcinoma cells. *Int J Mol Med*. (2014) 34:1430–8. doi: 10.3892/ijmm.2014.1925
33. Park JY, Juhnn YS. cAMP signaling increases histone deacetylase 8 expression via the Epac2-Rap1A-Akt pathway in H1299 lung cancer cells. *Exp Mol Med*. (2017) 49:e297. doi: 10.1038/emm.2016.152
34. Monteith GR, Prevarskaya N, Roberts-Thomson SJ. The calcium-cancer signalling nexus. *Nat Rev Cancer*. (2017) 17:367–80. doi: 10.1038/nrc.2017.18
35. Wang LL, Chen ZS, Zhou WD, Shu J, Wang XH, Jin R, et al. Down-regulated GATA-1 up-regulates interferon regulatory factor 3 in lung adenocarcinoma. *Sci Rep*. (2017) 7:2551. doi: 10.1038/s41598-017-02700-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 21 January 2020
doi: 10.3389/fonc.2019.01474

Comprehensive Analysis of Competitive Endogenous RNAs Network, Being Associated With Esophageal Squamous Cell Carcinoma and Its Emerging Role in Head and Neck Squamous Cell Carcinoma

Donghu Yu[1,2], *Xiaolan Ruan*[3], *Jingyu Huang*[4], *Weidong Hu*[4], *Chen Chen*[1,2], *Yu Xu*[5], *Jinxuan Hou*[6]* *and Sheng Li*[1,2]*

[1] Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan, China, [2] Human Genetics Resource Preservation Center of Hubei Province, Wuhan, China, [3] Department of Hematology, Renmin Hospital of Wuhan University, Wuhan, China, [4] Department of Thoracic Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China, [5] Department of Radiation and Medical Oncology, Zhongnan Hospital of Wuhan University, Wuhan, China, [6] Department of Thyroid and Breast Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China

Esophageal squamous cell carcinoma (ESCC) is a common malignancy with poor prognosis and survival rate. To identify meaningful long non-coding RNA (lncRNA), microRNA (miRNA), and messenger RNA (mRNA) modules related to the ESCC prognosis, The Cancer Genome Atlas-ESCC was downloaded and processed, and then, a weighted gene co-expression network analysis was applied to construct lncRNA co-expression networks, miRNA co-expression networks, and mRNA co-expression networks. Twenty-one hub lncRNAs, seven hub miRNAs, and eight hub mRNAs were clarified. Additionally, a competitive endogenous RNAs network was constructed, and the emerging role of the network involved in head and neck squamous cell carcinoma (HNSCC) was also analyzed using several webtools. The expression levels of eight hub genes (TBC1D2, ATP6V0E1, SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA) were different between normal esophageal tissues and HNSCC tissues. The expression levels of TBC1D2 and ATP6V0E1 were related to the survival time of HNSCC. The competitive endogenous RNAs network might provide common mechanisms involving in ESCC and HNSCC. More importantly, useful clues were provided for clinical treatments of both diseases based on novel molecular advances.

Keywords: esophageal squamous cell carcinoma, head and neck squamous cell carcinoma, prognosis, weighted gene co-expression network analysis, competitive endogenous RNAs network

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Mingjun Bi,
The University of Texas Health Science Center at San Antonio, United States
Xiaowen Chen,
Harbin Medical University, China

***Correspondence:**
Jinxuan Hou
jhou@whu.edu.cn
Sheng Li
lisheng-znyy@whu.edu.cn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *08 August 2019*
Accepted: *09 December 2019*
Published: *21 January 2020*

Citation:
Yu D, Ruan X, Huang J, Hu W, Chen C, Xu Y, Hou J and Li S (2020) Comprehensive Analysis of Competitive Endogenous RNAs Network, Being Associated With Esophageal Squamous Cell Carcinoma and Its Emerging Role in Head and Neck Squamous Cell Carcinoma. Front. Oncol. 9:1474. doi: 10.3389/fonc.2019.01474

INTRODUCTION

Esophageal squamous cell carcinoma (ESCC) is the globally predominant pathological type of esophageal cancer (1). For the lack of effective biomarkers, most patients with ESCC are diagnosed at a late stage, which leads to the poor prognosis of ESCC, with a 5-year survival rate of <20% (2, 3). Numerous studies have shown that T stage was the independent factor which influenced the prognosis of ESCC. Besides, most patients with ESCC have a high prevalence of second primary head and neck squamous cell carcinoma (HNSCC) (4). In Taiwan, 15–20% of patients with ESCC may develop a secondary HNSCC (5). Nowadays, it is necessary to do routine screening of head and neck field for the patients with newly diagnosed ESCC and that results in more frequent detection of second primary HNSCC. Therefore, it is of great value to identify the molecular mechanisms related to the development and the prognosis of ESCC, and further research for ESCC-HNSCC pathogenesis is also urgently needed.

Long non-coding RNA (lncRNA) refers to a non-coding RNA transcript with a length >200 nucleotides (6). In recent years, increasing evidences have revealed that multiple lncRNAs can play as potential biomarkers for the prognosis prediction of ESCC, including RNA-PCAT-1 (7), TTN-AS1 (8), and linc00460 (9). However, studies of single lncRNA cannot meet the requirement for exploration of ESCC prognosis. A lncRNA–microRNA (miRNA)–messenger RNA (mRNA) network, which is involved in many important cellular pathways, is badly needed to clarify exact mechanisms.

The competing endogenous RNA (ceRNA) hypothesis was presented by Salmena et al., which stated that mRNAs, lncRNAs, and other non-coding RNAs can act as natural miRNA "sponges" with common MREs to regulate the expression levels of certain genes (10). Nowadays, more and more studies have proven that the ceRNA regulation theory plays an important role in the development of cancer (11). For example, lncRNA-TTN-AS1 was identified to be a target of miR133b, and miR133b can repress the mRNA of fascin homolog 1 in ESCC. Further experiments demonstrated that lncRNA-TTN-AS1 could operate as a ceRNA for binding the microRNA to regulate the expression level of fascin homolog 1 (8).

Although Xue has reported differently expressed lncRNAs, miRNAs, and mRNAs between normal and ESCC tissues (12), the relationships between hub RNAs and important clinical traits had not been rigorously studied. To fulfill these gaps, mRNA co-expression networks, miRNA co-expression networks, and lncRNAs co-expression network were constructed by weighted gene co-expression network analysis (WGCNA) to identify mRNA, miRNA, and lncRNA modules related to T stage in ESCC. WGCNA is a method of mining module information from sequencing data. Under certain conditions, module is defined as a group of genes with similar expression changes in physiological process. This method seems similar to cluster analysis, and the difference is that WGCNA has a biological significance (13). The relationships between the modules and clinical features could be further explored to select candidate biomarkers for cancers. The relationships between lncRNAs and miRNAs, and miRNAs and mRNAs were predicted to build the lncRNA–miRNA–mRNA network, which would provide more information about the mechanisms of ESCC progression, even ESCC-HNSCC pathogenesis.

MATERIALS AND METHODS

Data Collection and Processing

A brief workflow for this study is shown in **Figure 1**. The RNA sequencing data of 95 samples with ESCC were retrieved from The Cancer Genome Atlas (TCGA) data portal (https://cancergenome.nih.gov/), which had been derived from the IlluminaHiSeq_RNASeq and the IlluminaHiSeq_miRNASeq sequencing platforms. Ninety-five samples were divided into two groups: 17 normal samples and 78 tumor samples. Gene expression profiles (GSE20437 and GSE38129) related to ESCC, which were downloaded for the validation from Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/), provided validation for selected hub mRNAs. The details of GSE20437 and GSE38129 are listed in **Table S1**. All datasets were normalized with quantile normalization. Analysis of variance were performed for TCGA-ESCC-mRNA and TCGA-ESCC-lncRNA. We chose the top 25% most variant mRNAs (4,938 mRNAs) and the top 25% most variant lncRNAs (3,712 genes) for constructing networks, while we did not do pretreatment for miRNA expression profile due to the small number of miRNAs (1,881 miRNAs).

Construction of Co-expression Networks

WGCNA was used to construct mRNA, miRNA, and lncRNA co-expression networks (14). The processes for constructing co-expression networks were similar. Thus, we took the construction of weighted mRNA co-expression networks as an example. First, a matrix of similarity was constructed by calculating the correlations of the processed genes. Then, an appropriate power of β was chosen as the soft-thresholding parameter to construct a scale-free network. Next, the adjacency was transformed into a topological overlap matrix (TOM) using TOM similarity, and the corresponding dissimilarity (1—TOM) was figured and the dissimilarity of module eigengenes (MEs) estimated. Last, the mRNAs with similar expression levels were categorized into the same module by DynamicTreeCut algorithm (15).

Identification of Clinically Significant Modules

The clinical trait we were concerned was T stage in ESCC patients and key modules which needed to be found in three networks separately. Above all, we worked out the relationship between clinical phenotype and MEs. MEs were deemed to represent the expression levels of all mRNAs, miRNAs, or lncRNAs in the related module. In addition, mediated *P*-value of each mRNA, miRNA, or lncRNA was calculated, and then, we worked out gene, miRNA, or lncRNA significance (GS = lg *P*). Finally, we selected the most clinically significant module according to module significance, which was the average GS of mRNAs, miRNAs, or lncRNAs involved in the related module. Besides, the connectivity of module was measured by absolute value of the

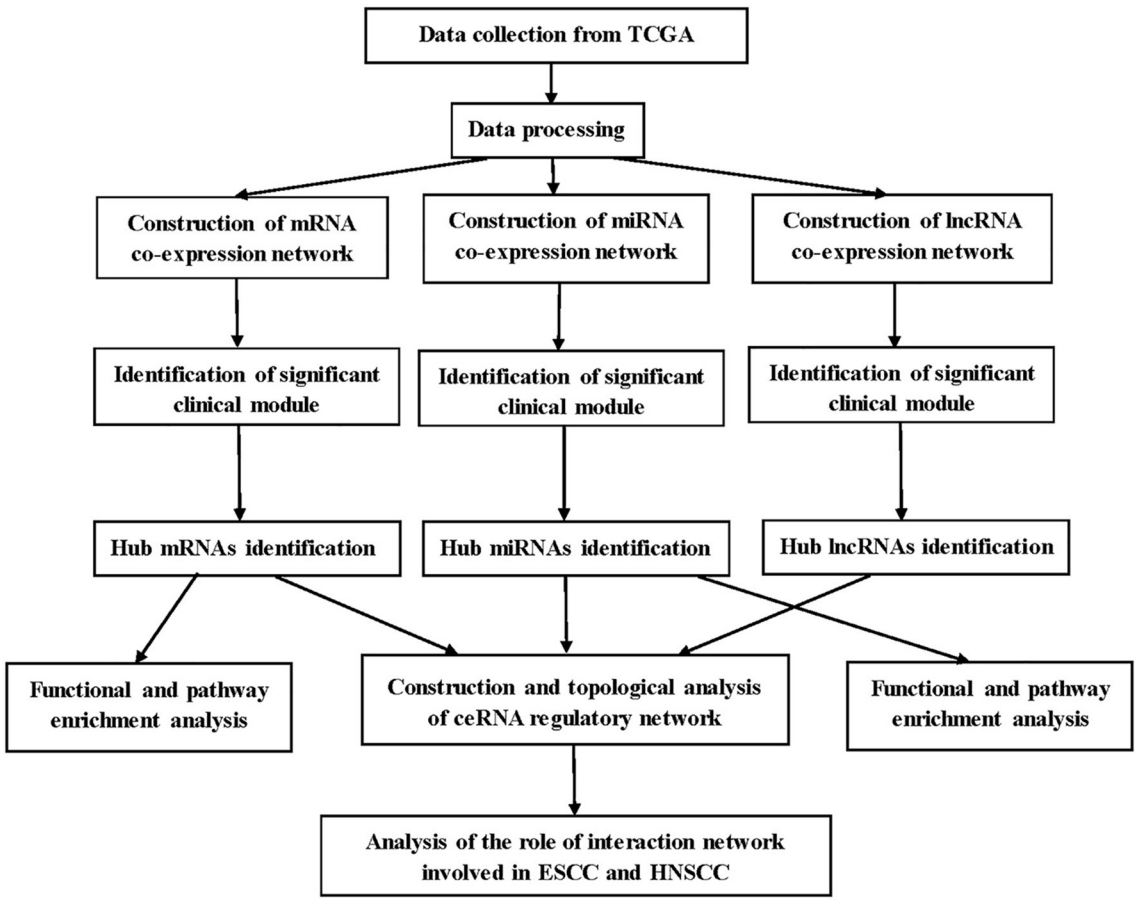


FIGURE 1 | Flow chart of data preparation, processing, and analysis.

Pearson's correlation, and the relationships between clinical trait and mRNAs, miRNAs, or lncRNAs were measured by absolute value of the Pearson's correlation. To build a ceRNA regulatory network in ESCC better, two modules in each co-expression network were selected. The RNA expression levels in one module were positively correlated with the clinical trait (T stage), and the RNA expression levels in the other module were negatively correlated with the T stage of ESCC.

Functional and Pathway Enrichment Analysis

The Database for Annotation, Visualization, and Integrate Discovery (DAVID) (https://david.ncifcrf.gov/) is a database for several kinds of functional annotation (16). With the help of Database for Annotation, Visualization, and Integrate Discovery, we identified biological meaning of the mRNAs in hub modules according to false discovery rate (FDR) < 0.05. Gene Ontology (GO) includes three terms: biological process (BP), cellular component (CC), and molecular function (MF); GO (BP, CC, MF) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses for the miRNAs in the hub modules were conducted using mirPath v.3, an online tool for miRNA pathway analysis (17). GO (BP, CC, MF) and KEGG enrichment analyses for the lncRNAs in the hub modules were conducted using co-lncRNA, a web-based computational tool that allows users to identify GO annotations and KEGG pathways that may be affected by co-expressed protein-coding genes of a single or multiple lncRNAs (18).

Identification and Validation of Hub mRNAs in ESCC

To identify real hub mRNAs associated with the development of ESCC, three methods were used to screen candidate mRNAs. First, the mRNAs that have high connectivity with module and selected phenotype were chosen as candidate genes in hub module [|cor. module membership| (|MM|) > 0.35]. Then, the protein/gene interactions for the mRNAs in each hub module were analyzed using STRING (19), and the mRNAs connected with more than four nodes in PPI network were selected as candidate mRNAs for further study. Next, survival analysis was performed for the mRNAs in each hub module by survival package in R, and the mRNAs with $P < 0.05$ were considered to be associated with overall survival in ESCC. Then, the common candidate mRNAs in three parts were considered as hub mRNAs.

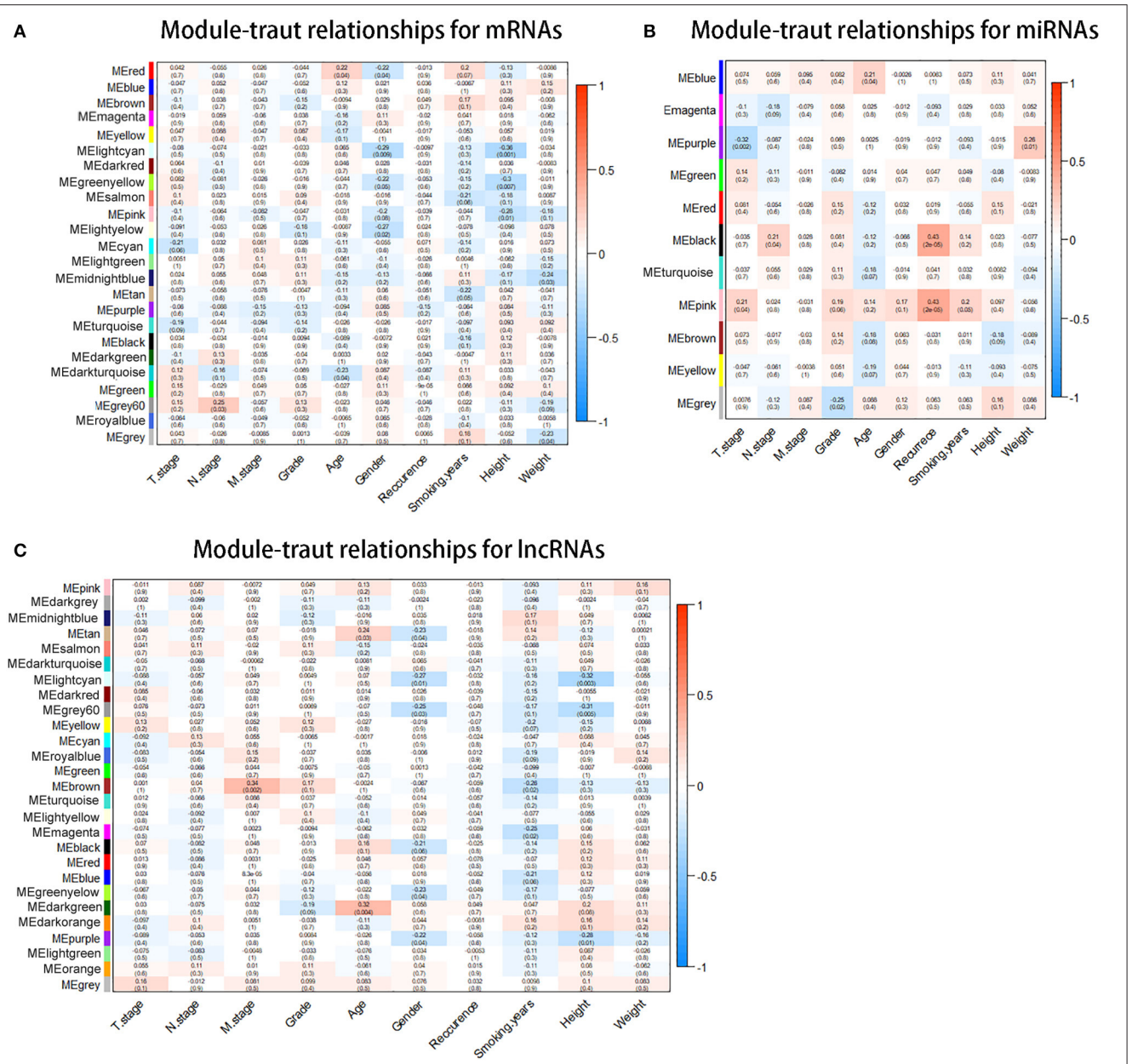


FIGURE 2 | Identification of modules associated with the clinical traits of esophageal squamous cell carcinoma (ESCC). **(A)** Distribution of average messenger RNA (mRNA) significance and errors in the modules associated with the T stage in ESCC. **(B)** Distribution of average microRNA (miRNA) significance and errors in the modules associated with the T stage. **(C)** Distribution of average long non-coding RNA (lncRNA) significance and errors in the modules associated with the T stage of esophageal squamous cell carcinoma (ESCC).

To verify our results, GSE20347 (including 17 normal esophageal tissues and 17 ESCC tissues) and GSE38129 (including 30 normal esophageal tissues and 30 ESCC tissues) were used to validate the different expression levels of hub mRNAs between normal tissues and ESCC tissues. Under the threshold of $|\log_2 FC| > 1.5$ and FDR < 0.05, differently expressed genes (DEGs) were selected by "limma" package in R in two datasets, separately. OSescc, containing survival data from GSE53625 and TCGA and giving users the ability to create publication-quality Kaplan–Meier plots (20), was used to further explore the prognostic biomarker in the dataset GSE53625 (21).

Identification Hub miRNAs and lncRNAs

The interactions between lncRNA and miRNA, and mRNA and miRNA could be predicted. As for selecting hub miRNAs, TargetScan (http://www.targetscan.org/) was employed to predict candidate miRNAs for hub mRNAs (22, 23), and context++ score of TargetScan > 0.4 were selected as threshold. Then, the

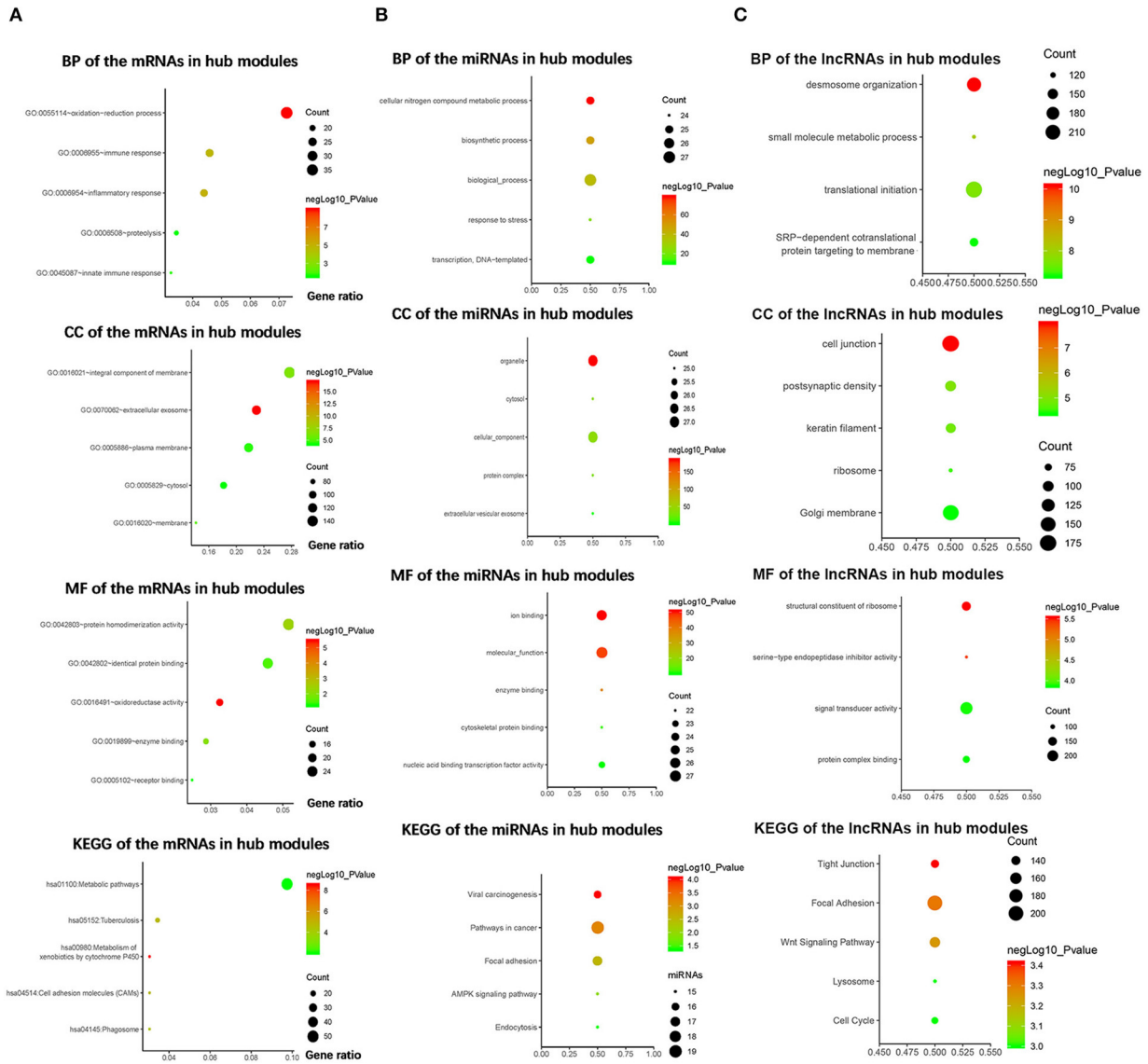


FIGURE 3 | Bioinformatics analysis of the messenger RNAs (mRNAs) and the microRNAs (miRNAs) in hub modules. **(A)** Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment of the mRNAs in green and cyan modules. **(B)** GO analysis and KEGG pathway enrichment of the miRNAs in pink and purple modules. **(C)** GO analysis and KEGG pathway enrichment of the lncRNAs in the yellow and midnight blue modules.

common candidate miRNAs with $|MM| > 0.4$ in hub modules and prediction by TargetScan was defined as real hub miRNAs. LncBase (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2) was used to predict lncRNA and miRNA interactions (24), and the score of LncBase > 0.7 was selected as threshold. The common candidate lncRNAs with $|MM| > 0.7$ in hub modules and prediction by LncBase were defined as real hub lncRNAs.

Construction and Topological Analysis of ceRNA Regulatory Network in ESCC

According to the prediction of TargetScan and LncBase, the interactions were used to construct the lncRNA–miRNA–mRNA network applying the Cytoscape software, and the interaction between genes was also demonstrated from STRING (25). It is well-known that hub nodes play critical roles in biological networks. Simultaneously, all node degrees of the lncRNA–miRNA–mRNA network were calculated by "NetworkAnalyzer" in Cytoscape.

The Prognostic Factors of ceRNA Network in ESCC and HNSCC

Survival analysis was performed for the mRNAs/miRNAs/lncRNAs in ceRNA network by survival package in R, and the threshold was selected as $P < 0.05$. In addition, to explore the role of the interaction network in HNSCC, UALCAN (http://ualcan.path.uab.edu/) was used to find the different expression levels of hub genes between normal

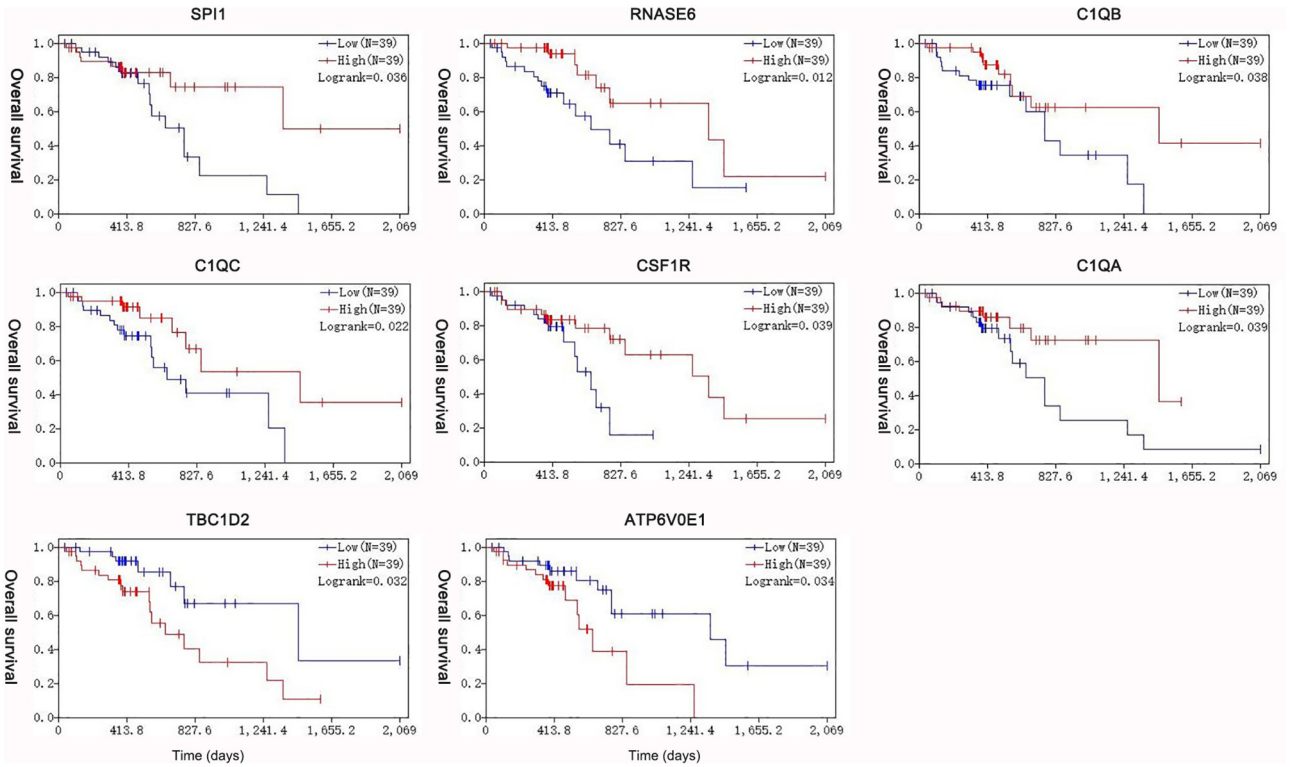


FIGURE 4 | Survival analysis of the association between the expression levels of hub mRNAs based on The Cancer Genome Atlas-esophageal squamous cell carcinoma (TCGA-ESCC). The expression levels of SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA were positively correlated with the overall survival. The expression levels of TBC1D2 and ATP6V0E1 were negatively correlated with the overall survival of ESCC.

tissues and cancer tissues. UALCAN is a useful online tool for analyzing cancer transcriptome data, which is based on public cancer transcriptome data (TCGA and MET500 transcriptome sequencing) (26). OncomiR (http://www.oncomir.org/), an online resource for exploring miRNA dysregulation in cancer based on TCGA, was used to find the different expression levels of hub miRNAs between normal tissues and cancer tissues (27). To explore the expression levels of hub lncRNAs in normal and HNSCC samples, independent *t*-test was performed for the hub lncRNAs with the dataset of TCGA-HNSCC-lncRNA. Besides, OncoLnc (http://www.oncolnc.org/), containing survival data from 21 cancer studies performed by TCGA and giving users the ability to create publication-quality Kaplan–Meier plots, was used to explore the relationship between the expression levels of hub mRNAs/miRNAs/lncRNAs and the survival time of HNSCC (28).

Functional Annotation of the Hub Genes

Gene Set Enrichment Analysis (GSEA) was performed for hub mRNAs in TCGA-ESCC (29). In TCGA-ESCC, according to the median expression of this hub gene, 119 cases were classified into high- and low-expression group (high group, $n = 60$; low group, $n = 59$). Gene size > 100, $|ES| > 0.6$, nominal $P < 0.05$, and FDR $< 25\%$ were chosen as the cutoff criteria. Besides, Spearman correlation analysis was performed to explore pairwise gene expression correlation for hub genes in TCGA-ESCC. We calculated correlation coefficient absolute values, and the top 300 hub genes were selected for functional enrichment analysis. Based on the results, the potential functions of each hub gene were predicted, and the method thus bore the name of "guilt of association" (30).

RESULTS

Weighted Co-expression Networks Construction and Key Modules Identification

With the method of average linkage hierarchical clustering, the samples of TCGA-ESCC were well clustered. To ensure a scale-free network, power of $\beta = 5$ (scale-free $R^2 = 0.949$) was selected as the soft-thresholding parameter for mRNA co-expression networks (**Figure S1A**). Power of $\beta = 3$ (scale-free $R^2 = 0.939$) was selected for miRNA co-expression networks (**Figure S1B**). Power of $\beta = 5$ (scale-free $R^2 = 0.935$) was selected for lncRNA co-expression networks (**Figure S1C**). The clustering dendrograms of the mRNAs (**Figure S2A**), miRNAs (**Figure S2B**), and lncRNAs (**Figure S2C**) were generated. By "WGCNA" package in R, the mRNAs, the miRNAs, and the lncRNAs, which had similar expression levels, were divided into modules to construct co-expression networks, separately.

TABLE 1 | The corresponding GS and MM of the hub mRNAs, hub miRNAs and lncRNAs in hub modules.

	ID	Module	GS	MM
mRNA	TBC1D2	Green	0.066802	0.396126
mRNA	ATP6V0E1	Green	0.068227	0.443309
mRNA	SPI1	Cyan	−0.26978	0.92313
mRNA	RNASE6	Cyan	−0.14359	0.902145
mRNA	C1QB	Cyan	−0.13492	0.84643
mRNA	C1QC	Cyan	−0.12612	0.835847
mRNA	CSF1R	Cyan	−0.16647	0.830564
mRNA	C1QA	Cyan	−0.15212	0.826306
miRNA	hsa-miR-515-5p	Pink	0.175606	0.747637
miRNA	hsa-miR-519e-5p	Pink	0.15501	0.49922
miRNA	hsa-miR-6769b-5p	Pink	0.147596	0.410585
miRNA	hsa-miR-519d-5p	Pink	0.013167	0.541514
miRNA	hsa-miR-4707-3p	Purple	−0.24545	0.7482
miRNA	hsa-miR-6756-5p	Purple	−0.28336	0.908069
miRNA	hsa-miR-650	Purple	−0.32189	0.94761
lncRNA	RP5-1029K10.2	Yellow	0.173204	0.816681
lncRNA	ETV5-AS1	Yellow	0.13791	0.78535
lncRNA	RP11-440L14.1	Yellow	0.131855	0.84422
lncRNA	RP5-1184F4.5	Yellow	0.131485	0.854835
lncRNA	AC226118.1	Yellow	0.111381	0.76165
lncRNA	RP3-470B24.5	Yellow	0.108569	0.870981
lncRNA	RP5-1125A11.7	Yellow	0.107931	0.741911
lncRNA	CTD-2023N9.1	Yellow	0.104444	0.900808
lncRNA	RP11-332H14.2	Yellow	0.09411	0.797046
lncRNA	XIST	Yellow	0.089522	0.401167
lncRNA	AC141928.1	Yellow	0.072768	0.73042
lncRNA	RP5-1054A22.4	Yellow	0.072657	0.868935
lncRNA	C1orf213	Yellow	0.066348	0.746537
lncRNA	PSMG3-AS1	Yellow	0.063239	0.798
lncRNA	AC016735.1	Yellow	0.056561	0.704278
lncRNA	RP11-2H3.6	Yellow	0.035953	0.826073
lncRNA	RP11-504P24.8	Yellow	0.023092	0.706706
lncRNA	CTD-3018O17.3	Yellow	0.009725	0.726373
lncRNA	LINC01355	Yellow	0.001995	0.730689
lncRNA	RP11-327F22.6	Midnight blue	−0.04346	0.733582
lncRNA	RP11-275I4.2	Midnight blue	−0.06852	0.711478

miRNA, microRNA; lncRNA, long non-coding RNA; GS, gene significance; MM, module membership.

In mRNA co-expression networks, green module (GS = 0.15; containing 279 mRNAs) and cyan module (GS = −0.21; containing 92 mRNAs) showed the highest correlation with T stage of ESCC (**Figure 2A**). In miRNA co-expression networks, pink module (GS = 0.21; containing 46 miRNAs) and purple module (GS = −0.32; containing 38 miRNAs) showed the highest correlation with T stage of ESCC (**Figure 2B**). In lncRNA co-expression networks, yellow module (GS = 0.13; containing 180 lncRNAs) and midnight blue module (GS = −0.11; containing 71 lncRNAs) showed the highest correlation with T stage of ESCC (**Figure 2C**). Six modules from three networks were picked for following analysis as the clinically significant modules.

Functional and Pathway Enrichment Analysis

To explore the biological functions of the mRNAs in hub modules, the mRNAs were categorized into BP, CC, and MF. The outcome of GO and KEGG enrichment of the mRNAs in green and cyan module is shown in **Figure 3A**. The mRNAs in BP were generally enriched in oxidation–reduction process, immune response, inflammatory response, proteolysis, and innate immune response; the mRNAs in CC were mainly focused on integral component of membrane, extracellular exosome, plasma membrane, cytosol, and membrane; the mRNAs in MF were significantly focused on protein homodimerization activity, identical protein binding, oxidoreductase activity, enzyme binding, and receptor binding. The top five significantly enriched pathways in green and cyan module were metabolic pathways, tuberculosis, metabolism of xenobiotics by cytochrome P450, cell adhesion molecules, and phagosome. Top enriched GO terms for the miRNAs in pink and purple modules were the following: biological process, cellular nitrogen compound metabolic process, biosynthetic process, transcription, DNA-templated and response to stress in BP; organelle, cellular component, cytosol, protein complex, and extracellular vesicular exosome in CC; and molecular function, ion binding, nucleic acid binding transcription factor activity, enzyme binding, and cytoskeletal protein binding in MF. The pathway analysis was also performed for the miRNAs in hub modules. The top five significantly enriched pathways were pathways in cancer, focal adhesion, viral carcinogenesis, AMPK signaling pathway, and endocytosis (**Figure 3B**). Top enriched GO terms for the lncRNAs in yellow and midnight blue modules were as follows: desmosome organization, small molecule metabolic process, translational initiation, signal-recognition particle-dependent co-translational protein targeting to membrane, and keratinocyte differentiation in BP; Golgi membrane, cell junction, postsynaptic density, keratin filament, and ribosome in CC; signal transducer activity, structural constituent of ribosome, protein complex binding, serine-type endopeptidase inhibitor activity, and metallopeptidase activity in MF. The pathway analysis was also performed for the lncRNAs in hub modules. The top five significantly enriched pathways were focal adhesion, Wnt signaling pathway, tight junction, cell cycle, and lysosome (**Figure 3C**).

Identification and Validation of Hub mRNAs in ESCC

Under the threshold of |MM| > 0.35, 103 mRNAs in cyan module and 17 mRNAs in green module were considered as candidate genes. Then, the relationship between mRNAs in each module was identified from STRING (**Figure S3**), and we calculated the connectivity degree of each node in PPI. Sixty mRNAs in green module and 148 mRNAs with degrees ≥4 were considered as candidate mRNAs because they interacted with more proteins. As for the survival analysis, 17 mRNAs in green module and 29 mRNAs in cyan module were identified to be related to the overall survival in ESCC. To identify the common mRNAs in three parts, we performed Venn diagram

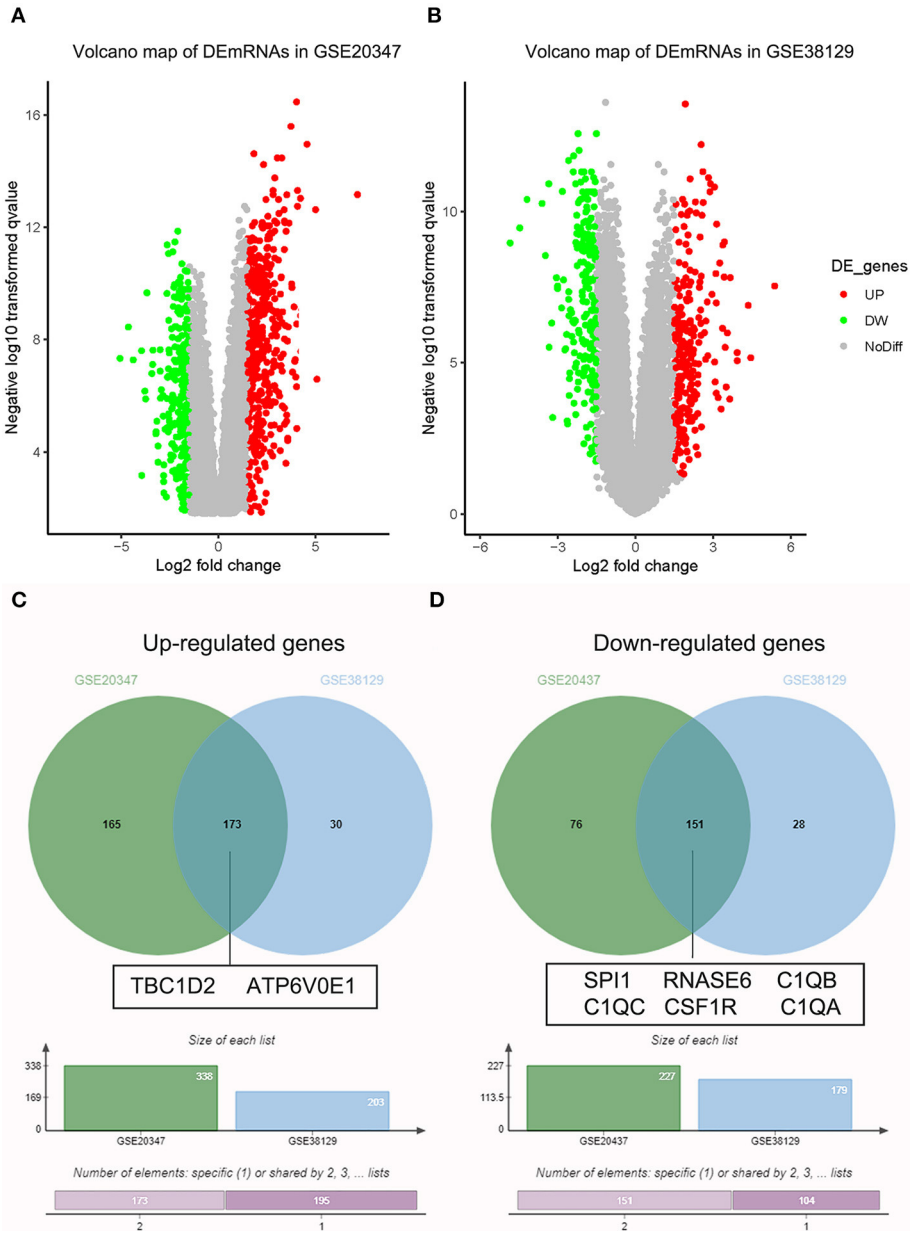


FIGURE 5 | Validation of hub messenger RNAs (mRNAs) in esophageal squamous cell carcinoma (ESCC). **(A)** Volcano plot visualizing differently expressed genes (DEGs) in GSE20347 (17 normal samples and 30 ESCC samples). **(B)** Volcano plot visualizing DEGs in GSE38129 (30 normal samples and 30 ESCC samples). **(C)** Identification of common upregulated genes between DEGs of GSE20347 and GSE38129. **(D)** Identification of common downregulated genes between DEGs of GSE20347 and GSE38129 by overlapping them.

by online tool jvenn (http://jvenn.toulouse.inra.fr/app/example.html) (**Figure S4**). Two mRNA (TBC1D2 and ATP6V0E1) in green module and six mRNAs (SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA) in cyan module were considered as real hub mRNAs, and they were closely related to the overall survival in ESCC (**Figure 4**). The corresponding MM and GS of the hub mRNAs in hub modules are shown in **Table 1**. GSE20347 and GSE38129 were used to validate the different expression levels of hub mRNAs between normal tissues and ESCC tissues with "limma" package in R. The results showed that TBC1D2 and ATP6V0E1 were significantly downregulated in ESCC ($\log_2$ FC $>$ 1.5 and FDR $<$ 0.05), while SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA are significantly downregulated ($\log_2$ FC $<$ -1.5 and FDR $<$ 0.05) (**Figure 5**). It is a pity that no other significant difference was observed in the prognostic analysis for the biomarkers in GSE53625 except for TBC1D2 (log-rank P = 0.028615) from OSescc.

TABLE 2 | The prediction of the interaction of hub mRNAs and hub miRNAs by Targetscan.

miRNA	Target gene	Context++ score of TargetScan
hsa-miR-519e-5p	RNASE6	−0.48
hsa-miR-515-5p	RNASE6	−0.48
hsa-miR-519d-5p	RNASE6	−0.45
hsa-miR-6756-5p	C1QA	−0.64
hsa-miR-6769b-5p	C1QA	−0.4
hsa-miR-4707-3p	TBC1D2	−0.59
hsa-miR-519d-5p	ATP6V0E1	−0.4
hsa-miR-650	ATP6V0E1	−0.59

mRNA, messenger RNA; miRNA, microRNA.

Identification of Hub miRNAs and lncRNAs

Based on the MM of miRNA co-expression network and the prediction by TargetScan (**Table 2**), seven miRNAs (hsa-miR-519e-5p, hsa-miR-519d-5p, hsa-miR-515-5p, hsa-miR-6756-5p, hsa-miR-6769b-5p, hsa-miR-4707-3p, and hsa-miR-650) were defined as real hub miRNAs. Based on the MM of lncRNA co-expression network and the prediction by LncBase (**Table 3**), 21 lncRNAs (RP11-275I4.2, RP11-327F22.6, LINC01355, CTD-3018O17.3, RP11-504P24.8, RP11-2H3.6, AC016735.1, PSMG3-AS1, C1orf213, RP5-1054A22.4, AC141928.1, XIST, RP11-332H14.2, CTD-2023N9.1, RP5-1125A11.7, RP3-470B24.5, AC226118.1, RP5-1184F4.5, RP11-440L14.1, ETV5-AS1, and RP5-1029K10.2) were considered as hub lncRNAs. The corresponding MM and GS of the hub miRNAs and the hub lncRNAs in hub modules are shown in **Table 1**.

Construction and Topological Analysis of ceRNA Regulatory Network in ESCC

Eight genes (SPI1, RNASE6, C1QB, C1QC, CSF1R, C1QA, TBC1D2, and ATP6V0E1), seven miRNAs (hsa-miR-519e-5p, hsa-miR-519d-5p, hsa-miR-515-5p, hsa-miR-6756-5p, hsa-miR-6769b-5p, hsa-miR-4707-3p, and hsa-miR-650), and 21 lncRNAs (RP11-275I4.2, RP11-327F22.6, LINC01355, CTD-3018O17.3, RP11-504P24.8, RP11-2H3.6, AC016735.1, PSMG3-AS1, C1orf213, RP5-1054A22.4, AC141928.1, XIST, RP11-332H14.2, CTD-2023N9.1, RP5-1125A11.7, RP3-470B24.5, AC226118.1, RP5-1184F4.5, RP11-440L14.1, ETV5-AS1, and RP5-1029K10.2) were involved in this interaction network. The lncRNA–miRNA–mRNA network is shown in **Figure 6A**. Besides, all node degrees of the network were calculated (**Table S2** and **Figure 6C**). According to the previous studies, a node with degree exceeding 5 was defined as a hub node (31, 32). In our study, eight nodes (including three mRNAs and five miRNAs) were selected as hub nodes. In addition, we calculated the number of the relationship pairs of miRNA–mRNA and lncRNA–miRNA, and the results are shown in **Table 4**. We found that three miRNAs (hsa-miR-519e-5p, hsa-miR-515-5p, and hsa-miR-6756-5p) not only had higher node degrees but also had a higher number of miRNA–mRNA and lncRNA–miRNA pairs. The results suggested that the miRNAs (hsa-miR-519e-5p, hsa-miR-515-5p, and hsa-miR-6756-5p) might play essential roles in ESCC progression, which would be considered as the key miRNAs.

TABLE 3 | The prediction of the interaction of hub lncRNAs and hub miRNAs by LncBase.

lncRNA	Target miRNA	The score of LncBase
XIST	hsa-miR-519e-5p	0.951
CTD-2023N9.1	hsa-miR-519e-5p	0.711
RP5-1184F4.5	hsa-miR-519e-5p	0.71
RP11-440L14.1	hsa-miR-519e-5p	0.987
RP11-332H14.2	hsa-miR-519e-5p	0.774
ETV5-AS1	hsa-miR-519e-5p	0.803
RP11-327F22.6	hsa-miR-519e-5p	0.706
AC141928.1	hsa-miR-519e-5p	0.707
AC016735.1	hsa-miR-519e-5p	0.779
RP5-1054A22.4	hsa-miR-519d-5p	0.726
RP11-327F22.6	hsa-miR-519d-5p	0.712
XIST	hsa-miR-515-5p	0.949
CTD-2023N9.1	hsa-miR-515-5p	0.711
RP5-1184F4.5	hsa-miR-515-5p	0.736
RP11-440L14.1	hsa-miR-515-5p	0.989
RP11-332H14.2	hsa-miR-515-5p	0.782
ETV5-AS1	hsa-miR-515-5p	0.767
AC141928.1	hsa-miR-515-5p	0.715
AC016735.1	hsa-miR-515-5p	0.797
XIST	hsa-miR-6756-5p	0.948
RP5-1029K10.2	hsa-miR-6756-5p	0.944
PSMG3-AS1	hsa-miR-6756-5p	0.711
AC226118.1	hsa-miR-6756-5p	0.7
C1orf213	hsa-miR-6756-5p	0.919
RP5-1125A11.7	hsa-miR-6756-5p	0.73
LINC01355	hsa-miR-6756-5p	0.773
RP11-2H3.6	hsa-miR-6769b-5p	0.919
AC226118.1	hsa-miR-6769b-5p	0.703
CTD-3018O17.3	hsa-miR-6769b-5p	0.848
RP11-275I4.2	hsa-miR-6769b-5p	0.991
RP11-440L14.1	hsa-miR-4707-3p	0.822
RP3-470B24.5	hsa-miR-650	0.815
C1orf213	hsa-miR-650	0.716
CTD-3018O17.3	hsa-miR-650	0.762
RP11-504P24.8	hsa-miR-650	0.974

lncRNA, long non-coding RNA; miRNA, microRNA.

The Prognostic Factors of ceRNA Network in ESCC and HNSCC

The R survival package was used for survival analysis for all RNAs in the ceRNA network. Because the overall survival of mRNAs was performed to select hub mRNAs ($P < 0.05$), the mRNAs in the ceRNA network were significantly associated with overall survival of ESCC. Through the Kaplan–Meier curve analysis for TCGA-ESCC, one miRNA (hsa-miR-515-5p) and one lncRNA (XIST) were found to be significantly

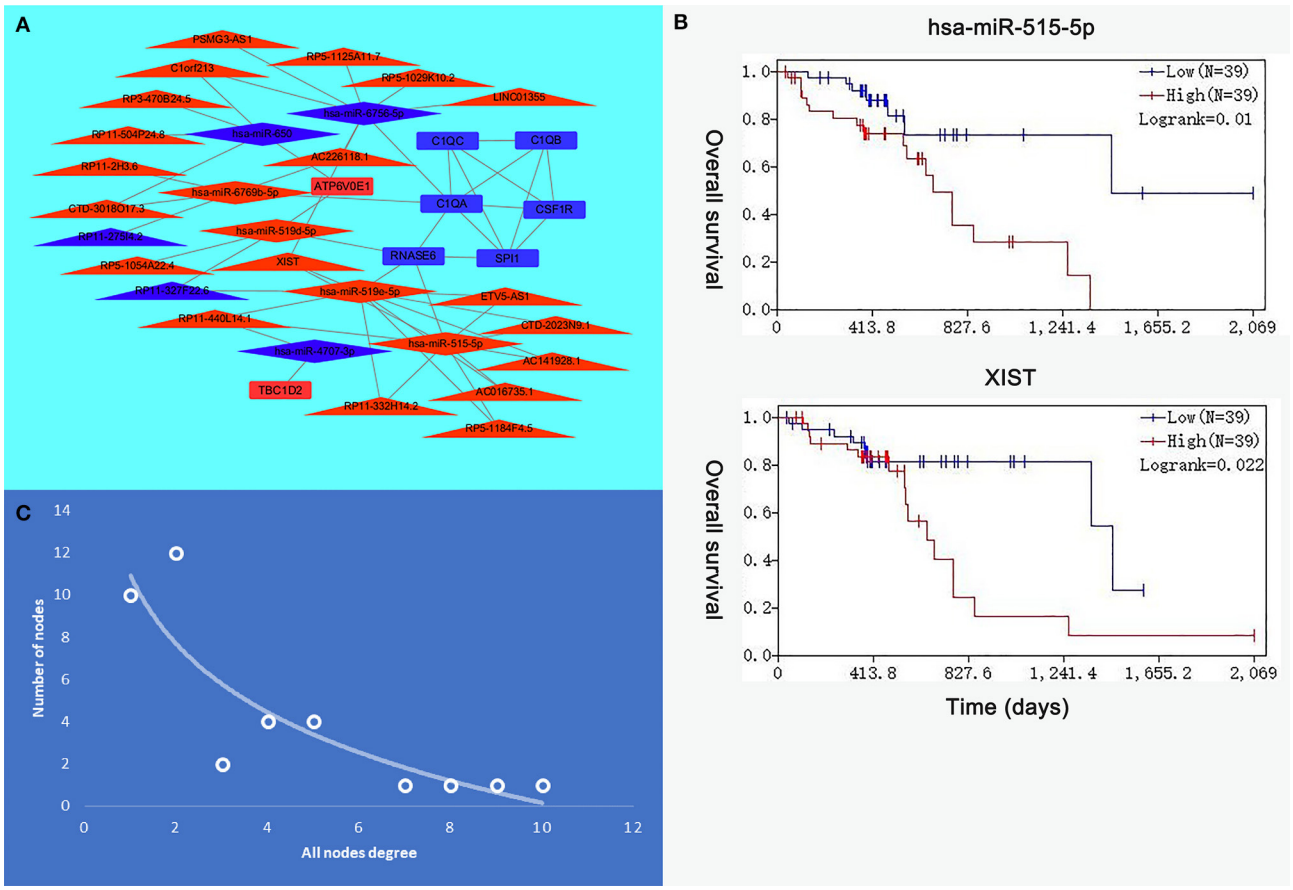


FIGURE 6 | The interaction network of hub microRNAs (miRNAs) and hub genes. **(A)** The view of the long non-coding RNA (lncRNA)–miRNA–messenger RNA (mRNA) network. The triangle represents lncRNAs, the rhombus represents miRNAs, and the rectangle represents mRNAs. **(B)** The expression levels of hsa-miR-515-5p and XIST were negatively correlated with the overall survival. **(C)** All node degree analysis reveals specific properties of the lncRNA–miRNA–mRNA network.

associated with overall survival. We found that the expression levels of the hsa-miR-515-5p miRNA and XIST lncRNA were negatively correlated with the overall survival rate ($P < 0.05$; **Figure 6B**). Besides, some databases were used to explore the role of the interaction network in HNSCC. The levels of eight genes (SPI1, RNASE6, C1QB, C1QC, CSF1R, C1QA, TBC1D2, and ATP6V0E1) expression were higher in tumor samples from UALCAN (**Figure 7A**). The results showed that the expression levels of the hub miRNAs/lncRNAs between normal and HNSCC tissues had no obvious difference. For the relationship between hub mRNAs/miRNAs/lncRNAs expression levels and the prognosis of HNSCC from OncoLnc, TBC1D2 and ATP6V0E1 negatively correlated with overall survival of HNSCC (**Figure 7B**). It is a pity that no other significant difference was observed in the prognostic analysis for the hub miRNAs/lncRNAs in HNSCC.

Functional Annotation of the Hub Genes

GSEA was performed to identify the lurking mechanisms related to ESCC progression of eight hub genes. As shown in **Table S3**, ESCC samples in TBC1D2 high-expression group were most significantly enriched in translational initiation molecules; ESCC samples in ATP6V0E1, SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA high-expression groups were most significantly enriched in adaptive immune response (**Tables S4–S10**). Based on the analysis of guilt of association, we identified that the hub genes were essential for T-cell activation, and they mainly played important roles in leukocyte cell–cell adhesion, regulation of lymphocyte activation, and T-cell receptor complex (**Figure S5**).

TABLE 4 | The number of lncRNA–miRNA and miRNA–mRNA pairs.

Number	Name	lncRNA–miRNA pairs	miRNA–mRNA pairs	Total number
1	hsa-miR-519e-5p	9	1	10
2	hsa-miR-515-5p	8	1	9
3	hsa-miR-6756-5p	7	1	8

lncRNA, long non-coding RNA; miRNA, microRNA.

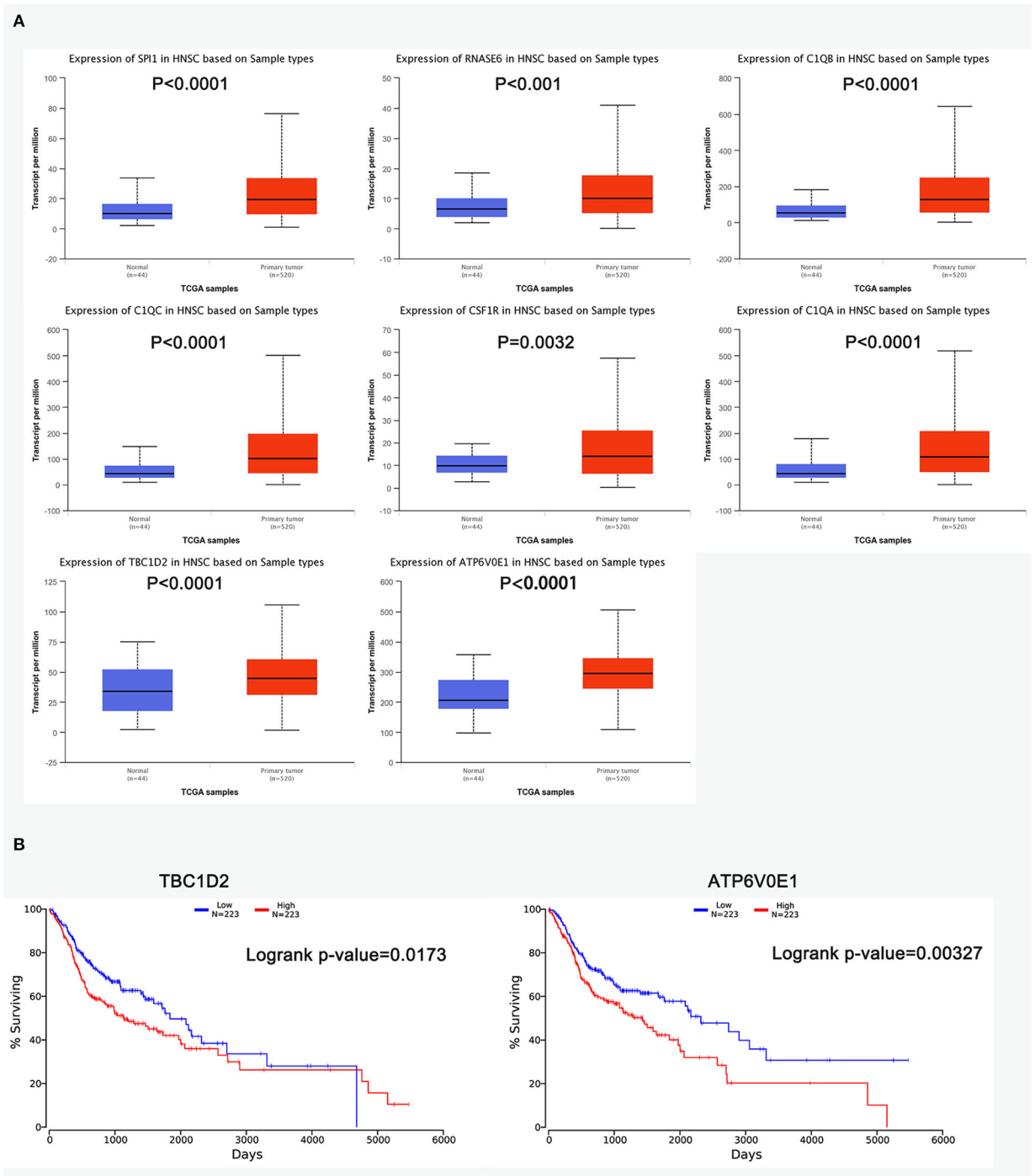


FIGURE 7 | The prognostic factors of competing endogenous RNA (ceRNA) network in head and neck squamous cell carcinoma (HNSCC). **(A)** Gene expression levels between normal and tumor samples [based on The Cancer Genome Atlas (TCGA)-HNSCC data in UALCAN]. **(B)** TBC1D2 and ATP6V0E1 were identified to be related to the overall survival of HNSCC from OncoLnc.

DISCUSSION

Although some certain chemotherapeutic drugs are used extensively for treating ESCC, including cisplatin (33, 34), docetaxel (33–35), nedaplatin (35), and fluorouracil (33–35), the prognosis of patients with ESCC is still very poor. Further development of some molecular drugs for ESCC is urgently required. In this study, it was the first time to identify ESCC mRNA, miRNA, and lncRNA modules by WGCNA at the same time. More importantly, the common mechanisms and molecular targets between ESCC and HNSCC were explored by bioinformatics analysis for the first time. We found six modules,

including two mRNA modules (green and cyan modules), two miRNA modules (pink and purple modules), and two lncRNA modules (yellow and midnight blue modules), which were significantly related to the T stage of ESCC. We identified eight hub mRNAs, seven hub miRNAs, and 21 hub lncRNAs, and the lncRNA–miRNA–mRNA network was constructed. Moreover, the drugs targeting the prognostic factors were collected from DrugBank (https://www.drugbank.ca/). Most of the prognostic factors were not used to develop targeting drugs yet, and more studies need to be done. Recently, Pexidartinib, a molecular drug targeting CSF1R, was approved by the Food and Drug Administration in August 2019 as the first systemic therapy for adult patients with symptomatic tensynovial giant cell tumor (36). This achievement would provide the reference to our latter work. In the independent validation of prognostic biomarkers in independent dataset, all of the samples of GSE53625 were collected in China, while the samples of TCGA-ESCC were collected in America. The predictive capability of the biomarkers in cancer patients prognosis will be changed greatly in different races (37, 38). We speculated the predication performance of these biomarkers for ESCC are different in different races. In the future, we will further explore these biomarkers for ESCC *in vivo* and *in vitro* and compare the predictability of the prognostic biomarkers from different ethnic groups with more precision experimental methods.

Previous studies have revealed that esophageal cancer stage was more important in predicting outcome of synchronous ESCC/HNSCC patients (5, 39). The lncRNA–miRNA–mRNA network, which was based on the RNA modules related to T stage of ESCC, would help us understand the pathogenesis of ESCC-HNSCC. In this study, TBC1D2 and ATP6V0E1 were identified to be related to the T stage of ESCC, and they have a significantly better chance of becoming molecular factors for the prognosis prediction in ESCC-HNSCC. The expression levels of TBC1D2 and ATP6V0E1 were increased in both ESCC and HNSCC tissues, and they are closely related to the overall survival of ESCC and HNSCC, which means that TBC1D2 and ATP6V0E1 could be common therapeutic targets for both cancers.

Most interestingly, we found that the expression levels of SPI1, RNASE6 C1QB, C1QC, CSF1R, and C1QA were downregulated in ESCC, whereas they were upregulated in HNSCC. Some certain genes patriciate different molecular mechanisms in different tumor cells, so the expression levels of the genes would be very different (40, 41). We speculated that these genes participate in different pathogenesis in ESCC and HNSCC, thus making significantly different expression levels of these genes in different cancers. Functional data about how these genes participating in ESCC and HNSCC are not enough, and further studies are needed to explore the proposed mechanism for this interesting phenomenon.

As for the miR-515-5p and XIST related to the survival of ESCC, we conducted a literature review of them. miR-515-5p was initially described as a placenta-specific factor participating in fetal growth (42). Previous studies have identified its important role in breast cancer and non-small cell lung cancer (43, 44). miR-515-5p overexpression could inhibit cell migration in both lung and breast cancers, which demonstrated that miR-515-5p could be a target of some molecular drugs treating the metastatic cancer patients (44). In this study, it is the first time to discover that the expression level of miR-515-5p is negatively related to the overall survival of ESCC, and miR-515-5p might control cancer cell progression through RNASE6 regulation. As for the lncRNA XIST (X-inactive specific transcript), it is the master regulator of X inactivation and a product of the XIST gene (45). More and more research indicates that lncRNA XIST plays an important role in cell proliferation and differentiation, and it is dysregulated in many cancers (46, 47). A recent study demonstrated the abnormal expression of XIST could contribute to esophageal cancer via miR-494/CDK6 axis (48). We found that XIST might influence the prognosis of ESCC via miR-6756-5p/C1QA. Functional data about how XIST participates in cancer pathology are not enough, and further studies are needed.

The mRNAs in the hub modules were generally enriched in oxidation–reduction process and immune response. Cancer cell survival depends on various redox-related mechanisms, which are targets of currently developed therapies (49). Besides, disruption of redox homeostasis is a crucial factor in the development of drug resistance for ESCC, which is a major problem facing current cancer treatment (50). The genes in the hub modules would help us better understand the new resistance mechanism of the drugs for ESCC, such as paclitaxel, fluorouracil, and cisplatin. The immune system has an important role in the control of tumor outgrowth. Nowadays, immunotherapy is a novel treatment option that has shown encouraging efficacy in several types of cancer, also in ESCC, and early phase evaluation of immune checkpoint inhibitors has yielded promising results (51). The genes, playing an important role in immune response, might be new targets for cancer immunotherapy. The miRNAs and the lncRNAs in the hub modules were generally enriched in cell division and cell adhesion. A lot of cancer-promoting errors may occur during cell division, such as DNA mutations and epigenetic mistakes, chromosome aberrations occurring, and the wrong distribution of cell-fate determinants between the daughter cells (52, 53). The miRNAs and the lncRNAs in the hub modules might regulate the enzyme genes relating to cell division to control tumor cells division and growth in ESCC. Cell adhesion molecules are involved in a series of important physiological and pathological processes, such as cell signal transduction and activation, cell extension and movement, and tumor metastasis (54). The expression levels of important cell adhesion molecules are of great significance for disease diagnosis, guiding clinical therapy, and prognosis in ESCC (55). For example, the high expression of EGFR causes the abnormal differentiation of ESCC cells and the decrease in adhesion between cells, and the tumor is prone to lymphatic and distant metastasis (56, 57).

This work not only identify the prognostic factors of ESCC but also do further research for ESCC-HNSCC pathogenesis. WGCNA, GO/KEGG analysis, GSEA, and some databases (UALCAN, OncomiR, and OncoLnc) were used to fully explore the common mechanisms involving in ESCC and HNSCC. Useful clues were provided for clinical treatment of both diseases based on novel molecular advances, but there are still insufficient exist. First, nowadays, many studies tried to identify genes associated

with progression and prognosis in patients with cancer using experimental methods. Lack of experiments (*in vivo* and *in vitro* validation) might be one limitation of our study. Second, the samples, suffering from ESCC and HNSCC, respectively, are not best one which is used to investigate mechanisms related to the prognosis of ESCC-HNSCC pathogenesis. We will further explore the ceRNA regulatory network and its role in the progression of ESCC-HNSCC using more in-depth bioinformatic analyses and experimental methods in the future.

In conclusion, the lncRNA–miRNA–mRNA network was conducted to explore the development of ESCC and common pathways between ESCC and HNSCC by WGCNA. We identified eight hub genes (TBC1D2, ATP6V0E1, SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA), one hub miRNA (hsa-miR-515-5p), and one lncRNA (XIST), which might be prognostic biomarkers for ESCC. In the future, the pathogenic overlap of ESCC and HNSCC may help us to clarify the common molecular mechanisms between both diseases and may provide a potential treatment strategy for both diseases.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://cancergenome.nih.gov/abouttcga/overview.

AUTHOR CONTRIBUTIONS

JHo and SL: conceived and designed the study. DY, XR, and JHu: performed the analysis procedures. DY, JHo, XR, CC, and YX: analyzed the results. WH and SL: contributed analysis tools. DY and JHo: contributed to the writing of the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by Zhongnan Hospital of Wuhan University Science, Technology and Innovation Cultivating Fund (znpy2018097) and the 351 Talent Project of Wuhan University (Luojia Young Scholars: SL) and The grant number of Young & Middle-aged Medical Key Talents Training Project of Wuhan is WHQG201901.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.01474/full#supplementary-material

Figure S1 | Determination of soft-thresholding power in the weighted gene co-expression network analysis (WGCNA). **(A)** Analysis of the scale-free fit index and the mean connectivity for various soft-thresholding powers for mRNA co-expression networks. **(B)** Analysis of the scale-free fit index and the mean connectivity for various soft-thresholding powers for miRNA co-expression networks. **(C)** Analysis of the scale-free fit index and the mean connectivity for various soft-thresholding powers for lncRNA co-expression networks.

Figure S2 | Clustering dendrograms. **(A)** Clustering dendrograms of the mRNAs based on a dissimilarity measure (1-TOM). **(B)** Clustering dendrograms of miRNAs based on a dissimilarity measure (1-TOM). **(C)** Clustering dendrograms of lncRNAs based on a dissimilarity measure (1-TOM).

Figure S3 | Protein-protein interaction networks for the genes in hub modules. **(A)** PPI network of 92 genes in cyan module. **(B)** PPI network of 279 genes in green module acquired from STRING 9.1.

Figure S4 | Identification of hub mRNAs in ESCC based on |MM| in co-expression networks, degrees in PPI network, and survival analysis. **(A)** TBC1D2 and ATP6V0E1 were considered as real hub mRNAs in green module. **(B)** SPI1, RNASE6, C1QB, C1QC, CSF1R, and C1QA were considered as real hub mRNAs in cyan module.

Figure S5 | Guilt of association for hub genes (SPI1, RNASE6, C1QB, C1QC, CSF1R, C1QA, TBC1D2, and ATP6V0E1).

Table S1 | Gene expression microarray datasets related to ESCC.

Table S2 | Node degree analysis for RNAs in ceRNA network.

Table S3 | Gene set enriched in esophageal samples with TBC1D2 high expression.

Table S4 | Gene set enriched in esophageal samples with ATP6V0E1 high expression.

Table S5 | Gene set enriched in esophageal samples with SPI1 low expression.

Table S6 | Gene set enriched in esophageal samples with RNASE6 low expression.

Table S7 | Gene set enriched in esophageal samples with C1QB low expression.

Table S8 | Gene set enriched in esophageal samples with C1QC low expression.

Table S9 | Gene set enriched in esophageal samples with CSF1R low expression.

Table S10 | Gene set enriched in esophageal samples with C1QA low expression.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
2. Yang W, Ma J, Zhou W, Zhou X, Cao B, Zhang H, et al. Molecular mechanisms and clinical implications of miRNAs in drug resistance of esophageal cancer. *Expert Rev Gastroenterol Hepatol.* (2017) 11:1151–63. doi: 10.1080/17474124.2017.1372189
3. Zhou J, Zhu J, Jiang G, Feng J, Wang Q. Downregulation of microRNA-4324 promotes the EMT of esophageal squamous-cell carcinoma cells via upregulating FAK. *Onco Targets Ther.* (2019) 12:4595–604. doi: 10.2147/OTT.S198333
4. Wang YK, Chuang YS, Wu TS, Lee KW, Wu CW, Wang HC, et al. Endoscopic screening for synchronous esophageal neoplasia among patients with incident head and neck cancer: Prevalence, risk factors, and outcomes. *Int J Cancer.* (2017) 141:1987–96. doi: 10.1002/ijc.30911
5. Chen Y-H, Lu H-I, Chien C-Y, Lo C-M, Wang Y-M, Chou S-Y, et al. Treatment outcomes of patients with locally advanced synchronous esophageal and head/neck squamous cell carcinoma receiving curative concurrent chemoradiotherapy. *Sci Rep.* (2017) 7:41785. doi: 10.1038/srep41785
6. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell.* (2009) 136:629–41. doi: 10.1016/j.cell.2009.02.006
7. Razavi M, Ghorbian S. Up-regulation of long non-coding rna-pcat-1 promotes invasion and metastasis in esophageal squamous cell carcinoma. *Excli J.* (2019) 18:422–8. doi: 10.17179/excli2018-1847
8. Lin C, Zhang S, Wang Y, Wang Y, Nice E, Guo C, et al. Functional role of a novel long noncoding RNA TTN-AS1 in esophageal squamous cell

carcinoma progression and metastasis. *Clin Cancer Res.* (2018) 24:486–98. doi: 10.1158/1078-0432.CCR-17-1851
9. Liang Y, Wu YY, Chen XD, Zhang SX, Wang K, Guan XY, et al. A novel long noncoding RNA linc00460 up-regulated by CBP/P300 promotes carcinogenesis in esophageal squamous cell carcinoma. *Biosci Rep.* (2017) 37:BSR20171019. doi: 10.1042/BSR20171019
10. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell.* (2011) 146:353–8. doi: 10.1016/j.cell.2011.07.014
11. Guo G, Kang Q, Zhu X, Chen Q, Wang X, Chen Y, et al. A long noncoding RNA critically regulates Bcr-Abl-mediated cellular transformation by acting as a competitive endogenous RNA. *Oncogene.* (2015) 34:1768–79. doi: 10.1038/onc.2014.131
12. Xue W-H, Fan Z-R, Li L-F, Lu J-L, Ma B-J, Kan Q-C, et al. Construction of an oesophageal cancer-specific ceRNA network based on miRNA, lncRNA, and mRNA expression data. *World J Gastroenterol.* (2018) 24:23–34. doi: 10.3748/wjg.v24.i1.23
13. Botia J, Vandrovcova J, Forabosco P, Hardy J, Lewis C, Ryten M, et al. An additional K-means clustering step improves the biological features of WGCNA gene co-expression networks. *Hum Hered.* (2015) 80:105–105. doi: 10.1186/s12918-017-0420-6
14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* (2008) 9:559. doi: 10.1186/1471-2105-9-559
15. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics.* (2008) 24:719–20. doi: 10.1093/bioinformatics/btm563
16. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* (2009) 4:44–57. doi: 10.1038/nprot.2008.211
17. Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* (2015) 43:W460–6. doi: 10.1093/nar/gkv403
18. Zhao Z, Bai J, Wu A, Wang Y, Zhang J, Wang Z, et al. Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database J Biol Databases Curation.* (2015) 2015:bav082. doi: 10.1093/database/bav082
19. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019) 47:D607–13. doi: 10.1093/nar/gky1131
20. Wang Q, Wang F, Lv J, Xin J, Xie L, Zhu W, et al. Interactive online consensus survival tool for esophageal squamous cell carcinoma prognosis analysis. *Oncol Lett.* (2019) 18:1199–206. doi: 10.3892/ol.2019.10440
21. Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, et al. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut.* (2014) 63:1700–10. doi: 10.1136/gutjnl-2013-305806
22. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* (2013) 41:W169–73. doi: 10.1093/nar/gkt393
23. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife.* (2015) 4:e05005. doi: 10.7554/eLife.05005.028
24. Paraskevopoulou MD, Vlachos IS, Karagkouni D, Georgakilas G, Kanellos I, Vergoulis T, et al. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* (2016) 44:D231–8. doi: 10.1093/nar/gkv1270
25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303
26. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia.* (2017) 19:649–58. doi: 10.1016/j.neo.2017.05.002
27. Wong NW, Chen YH, Chen S, Wang XW. OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics.* (2018) 34:713–5. doi: 10.1093/bioinformatics/btx627
28. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Comput Sci.* (2016) 2:e67. doi: 10.7717/peerj-cs.67
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
30. Yan X, Guo Z-X, Liu X-P, Feng Y-J, Zhao Y-J, Liu T-Z, et al. Four novel biomarkers for bladder cancer identified by weighted gene coexpression network analysis. *J Cell Physiol.* (2019) 234:19073–87. doi: 10.1002/jcp.28546
31. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature.* (2004) 430:88–93. doi: 10.1038/nature02555
32. Jiang H, Ma R, Zou SB, Wang YZ, Li ZQ, Li WP. Reconstruction and analysis of the lncRNA-miRNA-mRNA network based on competitive endogenous RNA reveal functional lncRNAs in rheumatoid arthritis. *Mol Biosyst.* (2017) 13:1182–92. doi: 10.1039/C7MB00094D
33. Zhu Y, Zhang W, Li Q, Li Q, Qiu B, Liu H, et al. A phase II randomized controlled trial: definitive concurrent chemoradiotherapy with docetaxel plus cisplatin versus 5-fluorouracil plus cisplatin in patients with oesophageal squamous cell carcinoma. *J Cancer.* (2017) 8:3657–66. doi: 10.7150/jca.20053
34. Okamoto H, Taniyama Y, Sakurai T, Heishi T, Teshima J, Sato C, et al. Definitive chemoradiotherapy with docetaxel, cisplatin, and 5-fluorouracil (DCF-R) for advanced cervical esophageal cancer. *Esophagus.* (2018) 15:281–5. doi: 10.1007/s10388-018-0627-7
35. Ohnuma H, Sato Y, Hayasaka N, Matsuno T, Fujita C, Sato M, et al. Neoadjuvant chemotherapy with docetaxel, nedaplatin, and fluorouracil for resectable esophageal cancer: a phase II study. *Cancer Sci.* (2018) 109:3554–63. doi: 10.1111/cas.13772
36. Tap WD, Gelderblom H, Palmerini E, Desai J, Bauer S, Blay J-Y, et al. Pexidartinib versus placebo for advanced tenosynovial giant cell tumour (ENLIVEN): a randomised phase 3 trial. *Lancet.* (2019) 394:478–87. doi: 10.1016/S0140-6736(19)30764-0
37. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. *JAMA.* (2015) 313:165–73. doi: 10.1001/jama.2014.17322
38. Ashktorab H, Kupfer SS, Brim H, Carethers JM. Racial disparity in gastrointestinal cancer risk. *Gastroenterology.* (2017) 153:910–23. doi: 10.1053/j.gastro.2017.08.018
39. Shinoto M, Shioyama Y, Sasaki T, Nakamura K, Ohura H, Toh Y, et al. Clinical results of definitive chemoradiotherapy for patients with synchronous head and neck squamous cell carcinoma and esophageal cancer. *Am J Clin Oncol.* (2011) 34:362–6. doi: 10.1097/COC.0b013e3181e84b4b
40. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature.* (2000) 406:747–52. doi: 10.1038/35021093
41. Cheng J, Guo Y, Gao Q, Li H, Yan H, Li M, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget.* (2017) 8:30265–75. doi: 10.18632/oncotarget.15754
42. Higashijima A, Miura K, Mishima H, Kinoshita A, Jo O, Abe S, et al. Characterization of placenta-specific microRNAs in fetal growth restriction pregnancy. *Prenat Diagn.* (2013) 33:214–22. doi: 10.1002/pd.4045
43. Pinho FG, Frampton AE, Nunes J, Krell J, Alshaker H, Jacob J, et al. Downregulation of microRNA-515-5p by the estrogen receptor modulates sphingosine kinase 1 and breast cancer cell proliferation. *Cancer Res.* (2013) 73:5936–48. doi: 10.1158/0008-5472.CAN-13-0158
44. Pardo OE, Castellano L, Munro CE, Hu Y, Mauri F, Krell J, et al. miR-515-5p controls cancer cell migration through MARK4 regulation. *EMBO Rep.* (2016) 17:570–84. doi: 10.15252/embr.201540970
45. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human x-inactivation center is expressed exclusively from the inactive x-chromosome. *Nature.* (1991) 349:38–44. doi: 10.1038/349038a0
46. Chen D-L, Ju H-Q, Lu Y-X, Chen L-Z, Zeng Z-L, Zhang D-S, et al. Long non-coding RNA XIST regulates gastric cancer progression by acting as a

molecular sponge of miR-101 to modulate EZH2 expression. *J Exp Clin Cancer Res*. (2016) 35:142. doi: 10.1186/s13046-016-0420-1
47. Yan X, Liu XP, Guo ZX, Tong-Zu L, Li S. Identification of hub genes associated with progression and prognosis in patients with bladder cancer. *Front Genet*. (2019) 10:408. doi: 10.3389/fgene.2019.00408
48. Chen ZZ, Hu X, Wu Y, Cong L, He X, Lu JW, et al. Long non-coding RNA XIST promotes the development of esophageal cancer by sponging miR-494 to regulate CDK6 expression. *Biomed Pharmacother*. (2019) 109:2228–36. doi: 10.1016/j.biopha.2018.11.049
49. Stevens A-S, Pirotte N, Wouters A, Van Roten A, Van Belleghem F, Willems M, et al. Redox-Related Mechanisms to Rebalance Cancer-Deregulated Cell Growth. *Curr Drug Targets*. (2016) 17:1414–37. doi: 10.2174/1389450116666150506112817
50. Yang W, Zou L, Huang C, Lei Y. Redox regulation of cancer metastasis: molecular signaling and therapeutic opportunities. *Drug Dev Res*. (2014) 75:331–41. doi: 10.1002/ddr.21216
51. Mimura K, Yamada L, Ujiie D, Hayase S, Tada T, Hanayama H, et al. Immunotherapy for esophageal squamous cell carcinoma: a review. *Fukushima J Med Sci*. (2018) 64:46–53. doi: 10.5387/fms.2018-09
52. Rad E, Tee AR. Neurofibromatosis type 1: fundamental insights into cell signalling and cancer. *Semin Cell Dev Biol*. (2016) 52:39–46. doi: 10.1016/j.semcdb.2016.02.007
53. Roehrich M, Koelsche C, Schrimpf D, Capper D, Sahm F, Kratz A, et al. Methylation-based classification of benign and malignant peripheral nerve sheath tumors. *Acta Neuropathol*. (2016) 131:877–87. doi: 10.1007/s00401-016-1540-6
54. Tang H, Jiang L, Zhu C, Liu R, Wu Y, Yan Q, et al. Loss of cell adhesion molecule L1 like promotes tumor growth and metastasis in esophageal squamous cell carcinomay. *Oncogene*. (2019) 38:3119–33. doi: 10.1038/s41388-018-0648-7
55. Natsuizaka M, Whelan KA, Kagawa S, Tanaka K, Giroux V, Chandramouleeswaran PM, et al. Interplay between Notch1 and Notch3 promotes EMT and tumor initiation in squamous cell carcinoma. *Nat Commun*. (2017) 8:1758. doi: 10.1038/s41467-017-01500-9
56. Yoshioka M, Ohashi S, Ida T, Nakai Y, Kikuchi O, Amanuma Y, et al. Distinct effects of EGFR inhibitors on epithelial- and mesenchymal-like esophageal squamous cell carcinoma cells. *J Exp Clin Cancer Res*. (2017) 36:101. doi: 10.1186/s13046-017-0572-7
57. Zhang W, Hong R, Xue L, Ou Y, Liu X, Zhao Z, et al. Piccolo mediates EGFR signaling and acts as a prognostic biomarker in esophageal squamous cell carcinoma. *Oncogene*. (2017) 36:3890–902. doi: 10.1038/onc.2017.15

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fbioe.2020.00012

Prioritizing Gene Cascading Paths to Model Colorectal Cancer Through Engineered Organoids

Yanyan Ping[1†], Chaohan Xu[1†], Liwen Xu[1†], Gaoming Liao[1†], Yao Zhou[1], Chunyu Deng[1], Yujia Lan[1], Fulong Yu[1], Jian Shi[1], Li Wang[1], Yun Xiao[1,2*] and Xia Li[1,2*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] Key Laboratory of Cardiovascular Medicine Research, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Deli Liu,
Weill Cornell Medicine, Cornell University, United States
Wei-Hua Chen,
Huazhong University of Science and Technology, China
Dijun Chen,
Nanjing University, China

***Correspondence:**
Xia Li
lixia@hrbmu.edu.cn
Yun Xiao
xiaoyun@ems.hrbmu.edu.cn
Li Wang
wangli@hrbmu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 12 October 2019
Accepted: 08 January 2020
Published: 04 February 2020

Citation:
Ping Y, Xu C, Xu L, Liao G, Zhou Y, Deng C, Lan Y, Yu F, Shi J, Wang L, Xiao Y and Li X (2020) Prioritizing Gene Cascading Paths to Model Colorectal Cancer Through Engineered Organoids. Front. Bioeng. Biotechnol. 8:12. doi: 10.3389/fbioe.2020.00012

Engineered organoids by sequential introduction of key mutations could help modeling the dynamic cancer progression. However, it remains difficult to determine gene paths which were sufficient to capture cancer behaviors and to broadly explain cancer mechanisms. Here, as a case study of colorectal cancer (CRC), functional and dynamic characterizations of five types of engineered organoids with different mutation combinations of five driver genes (*APC*, *SMAD4*, *KRAS*, *TP53,* and *PIK3CA*) showed that sequential introductions of all five driver mutations could induce enhanced activation of more hallmark signatures, tending to cancer. Comparative analysis of engineered organoids and corresponding CRC tissues revealed sequential introduction of key mutations could continually shorten the biological distance from engineered organoids to CRC tissues. Nevertheless, there still existed substantial biological gaps between the engineered organoid even with five key mutations and CRC samples. Thus, we proposed an integrative strategy to prioritize gene cascading paths for shrinking biological gaps between engineered organoids and CRC tissues. Our results not only recapitulated the well-known adenoma–carcinoma sequence model (e.g., AKST-organoid with driver mutations in *APC*, *KRAS*, *SMAD4*, and *TP53*), but also provided potential paths for delineating alternative pathogenesis underlying CRC populations (e.g., A-organoid with *APC* mutation). Our strategy also can be applied to both organoids with more mutations and other cancers, which can improve and innovate mechanism across cancer patients for drug design and cancer therapy.

Keywords: gene cascading paths, prioritizing, colorectal cancer, engineered organoids, random walk with restart

INTRODUCTION

The well-known adenoma–carcinoma sequence model described a basic carcinogenesis mechanism of colorectal cancer (CRC) (Vogelstein and Kinzler, 2004; Brenner et al., 2014). The sequential genetic alterations of *APC*, *KRAS*, *SMAD4,* and *TP53* could recapitulate the key features in transition from normal to adenoma and to initiation and progression of CRC, which promoted the understanding of pathogenesis in CRCs (Powell et al., 1992; Drost et al., 2015; Chen et al., 2016). Mutations on these genes could deregulate driver pathways to confer selective growth advantages and further to drive colorectal carcinogenesis. Tumor suppressor gene *APC* acted as an antagonist of the WNT signaling pathway. The inactivating mutations of *APC* could initiate a benign adenoma

by activating the WNT pathway (Powell et al., 1992; Roper et al., 2017; Takeda et al., 2019), which was proved by the upregulation of β-catenin driven by APC mutations (Matano et al., 2015). The follow genetic alterations in *KRAS*, *SMAD4,* and *TP53* further promoted the transition of adenoma to CRC by activating EGFR, P53 and TGF-β pathways (Drost et al., 2015; Chen et al., 2016). *KRAS* was reported to play driver roles during the progression from early to intermediate adenoma stages (Takeda et al., 2019). The activating mutations in *KRAS* could activate EGF signaling. The *SMAD4* and *TP53* mutations promoted the transition from adenoma to adenocarcinoma stages (Fearon and Vogelstein, 1990). *SMAD* mutations reduced the *SMAD* protein and inhibited TGF-β signaling pathway. The mutation in *TP53* could overexpressed a truncated *TP53* protein which made *TP53* lose tumor suppressor roles (Tang et al., 2019). However, due to the high heterogeneity of genetic alterations across CRC population, it was inefficient for these driver mutations to characterize the molecular mechanism of broad CRC patients. Prioritizing different gene cascading paths for directing sequential introduction of key mutations were the pressing problem.

Organoids, as an *in vitro* 3D models, could closely recapitulate genetic spectra of original tissues (Morizane et al., 2015). For example, tumor organoids closely recapitulated the molecular spectra in CRC (van de Wetering et al., 2015). Introducing key mutations into organoids other than cells could provide better manners to examine the influence of driver genes during cancer carcinogenesis. Directly targeting modification of cancer genes could produce cancer cells from the mouse primary cells or *in vivo* tissue (Ran et al., 2013; Heckl et al., 2014; Platt et al., 2014; Sánchez-Rivera et al., 2014; Xue et al., 2014). Driver gene-targeted engineered organoids could grow in hostile medium while normal intestinal organoids ceased proliferation. We summarized the recent studies modeling CRC using intestinal organoids with introducing driver mutations in *APC*, *SMAD4*, *KRAS*, *TP53,* and *PIK3CA* (**Table S1**) (Cooks et al., 2013; Onuma et al., 2013; Drost et al., 2015; Matano et al., 2015; Chen et al., 2016; Nakayama et al., 2017; O'Rourke et al., 2017; Riemer et al., 2017; van Lidth de Jeude et al., 2017). *APC* mutations activated WNT signaling and promoted the growth of intestinal organoids in medium lacking WNT signaling (Matano et al., 2015). Intestinal organoids with *APC* mutations developed into benign tumors after transplantation (O'Rourke et al., 2017). *SMAD4* mutation-targeted organoids could grew in condition without inhibitor of TGF-β receptor signaling that was essential for sustaining the growth of normal intestinal cells (Matano et al., 2015). Engineered organoids expressing *KRAS* mutations could expand in the condition withdrawing EGFR signaling (Matano et al., 2015). *TP53* mutations induced prolongation of activation of NF-kappaB signaling, and promoted inflammation-associated colorectal cancer (Cooks et al., 2013). *TP53* mutation-targeted organoids could recover in the condition of activation of *TP53* signaling pathway which can induce cell cycle arrest and apoptosis (Matano et al., 2015). Oncogenic *PIK3CA* could regulate cell motility though *AKT*, and *PIK3CA* mutations played key roles in reprograming glutamine metabolism in colorectal cancers (Hao et al., 2016). *PIK3CA* mutations could induce cell attachment and motility under cooperation of *CTNNB1* (Riemer et al., 2017). Oncogenic *PIK3CA* could regulate cell motility though *AKT*, and *PIK3CA* mutations played key roles in reprograming glutamine metabolism in colorectal cancers (Hao et al., 2016). Sequential introducing different combinations of these driver mutations could delineate the progression from normal epithelium to adenoma and carcinoma. Engineered organoids with *APC* and *KRAS* mutations grew into lager dysplasia without invasive features (Takeda et al., 2019), and further formed invasive submucosal tumor under condition of inhibited TGF-β signaling pathway (Chen et al., 2016; Takeda et al., 2019). These studies implied that engineered organoids with sequential introducing driver mutations could provide new clues to exploring developmental mechanisms of cancers. However, whether these engineered organoids were sufficient to capture broad cancer behaviors were still a challenge.

The transformation of normal cells to tumor cells was the dynamic dysregulated procession of cellular homeostasis, which was the requirement for the organism function normally (Rosenfeldt et al., 2013). The activity of biological functions could reflect the extent of homeostasis. Many functional activity-based methods were proposed to reveal the disease mechanisms (Lee et al., 2008; Gatza et al., 2010; Drier et al., 2013). The patterns of functional activity made tumor disease classification more precise and built subtype characterizations (Lee et al., 2008; Gatza et al., 2010). The function dysregulated scores characterized the deregulated extent of functions in individual samples (Drier et al., 2013). Measuring the difference of function activity among different cancer stages could help characterizing the dynamic progression of CRC.

In this work, from the single-mutant to quintuple-mutant engineered organoids, we dynamically characterized the function activities of hallmark signatures and measured the biological gaps between the engineered organoids and the CRC samples. An integrative strategy was designed to prioritize the gene cascading paths which could help us to understand the carcinogenesis mechanism of broad CRC patients with different profile of genetic alterations (**Figure 1**).

MATERIALS AND METHODS

Data Collection and Processing

Gene Expression Profiles and Mutation Profiles of Colorectal Cancer

We downloaded the gene expression profiles (GSE57965) of adenoma and engineered organoids (**Table S3**), which contained five adenoma samples with *APC* mutation (A-organoid), 1 adenoma sample with genetic modification of *SMAD4* deletion (AS-organoid), 1 adenoma sample of genetic modification of knocking in $KRAS^{G12V}$ (AK-organoid), 2 engineered human colon organoids carrying four gene mutations (*APC*, $KRAS^{G12V}$, *SMAD4*, and *TP53*, AKST-organoids) and 1 engineered human colon organoids carrying five gene mutations (*APC*, $KRAS^{G12V}$, *SMAD4*, *TP53,* and $PIK3CA^{E545K}$, AKSTP-organoid) (Matano et al., 2015). The gene expression profile with 20,014 genes were obtained after removing probes corresponding to multiple

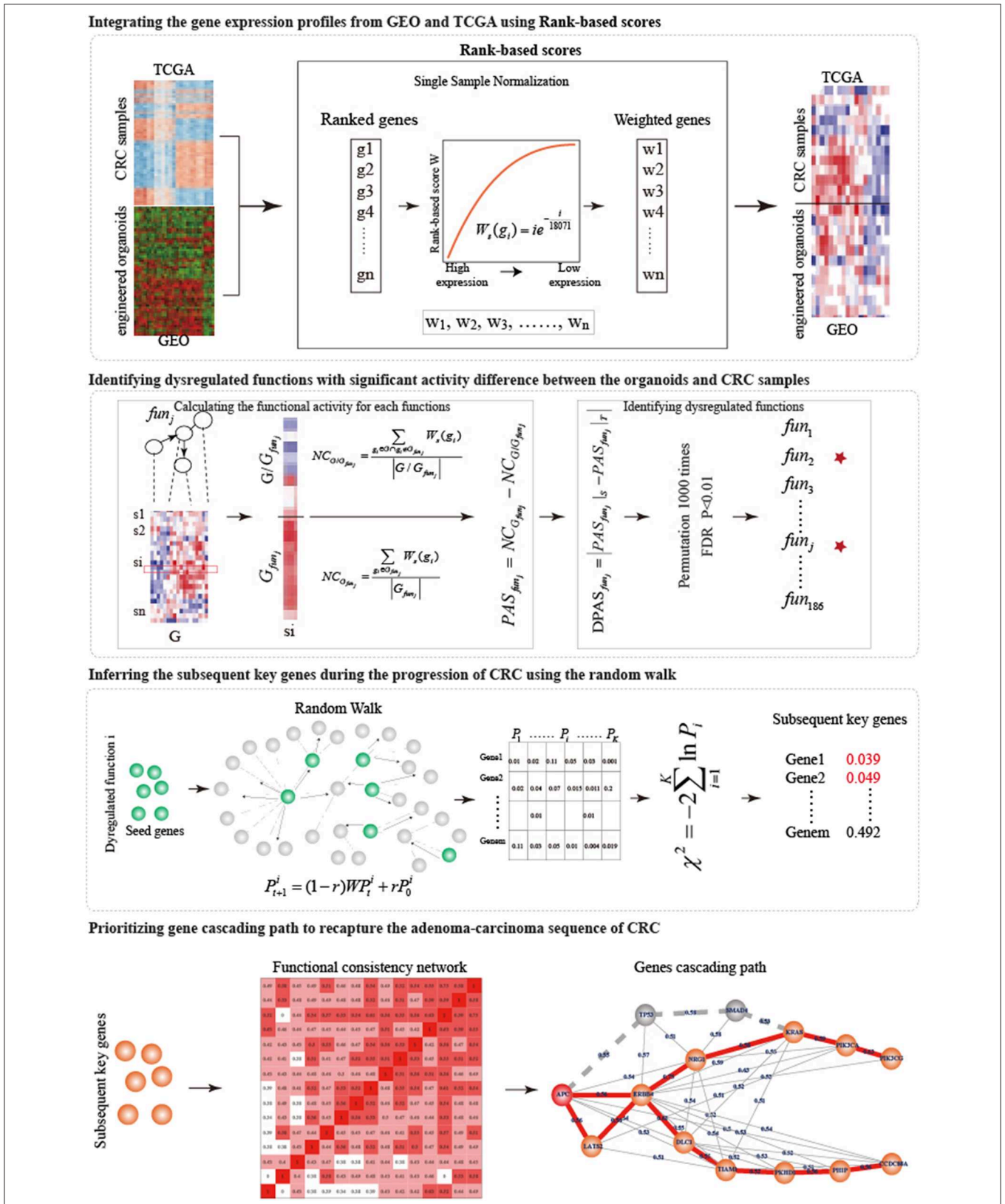


FIGURE 1 | The overview of the integrative strategy for prioritizing gene cascading path.

genes and averaging the expression level of multiple probes of each gene.

We also downloaded the somatic mutation data (level 2) and gene expression profiles (RNA-seq) of colorectal cancer from the cancer genome atlas (TCGA). We extracted a mutation profiles which contained the samples with mutations in at least one of five genes (including *APC*, *SMAD4*, *TP53*, *KRAS,* and *PIK3CA*) and removed mutation types of silent, intron and 5'UTR. Finally, we obtained 103 samples with both gene expression profile and mutation profile (**Table S3**), in which 54 samples only with APC mutation, 40 samples only with mutations in both *APC* and *KRAS*, 3 samples with mutations only in both APC and *SMAD4,* 1 sample with mutations only in four genes (*APC*, *KRAS*, *SMAD4,* and *TP53*), and five samples with mutations of all of five genes.

KEGG Pathways and HPRD Protein Interaction Network

We downloaded the KGMLs of 222 human pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). To get the topological information of these pathways, we got the corresponding undirected graphs of pathways and the degrees of genes in these pathways using the R package iSubpathwayMiner (Li et al., 2009). Only the pathways in which genes were connected with each other were kept. Finally, we obtained 186 pathways as the functions to characterize the biological gaps between organoids and cancer samples.

The protein interaction network was obtained from the Human Protein Reference Database (HPRD, version 9) (Keshava Prasad et al., 2009), which contained 9,617 genes and 39,240 interactions among these genes.

Methods

We proposed an integrative strategy to prioritize the gene cascading path for directing CRISPR-Cas9 to construct colorectal cancer organoids (**Figure 1**).

Integrating the Gene Expression Profiles From GEO and TCGA Using Rank-Based Scores

To joint analysis of expression profiles from GEO and TCGA, we used the Rank-based scores (Amar et al., 2015) to normalize the expression profiles of engineered organoids and CRC samples. 18,071 common genes were detected by both GEO and TCGA. For each sample *s*, the expression values of 18,071 genes were sorted in the decreasing order. Rank of highest expressed gene was 1 and that of lowest one was 18,071. The rank i of gene g was transformed into rank-based score: $W_s(g_i) = ie^{-\frac{i}{18071}}$. The rank-based scores of genes in the samples were used to joint analysis.

Identifying Dysregulated Functions in Biological Gaps Between Engineered Organoids and Corresponding CRC Samples

To investigate the potential driver capability of driver mutations, we characterized the biological distance from engineered organoids to CRC samples by identifying the dysregulated functions.

Functional Activity

Functional activity could measure the active status of biological functions in a specific sample (Bild et al., 2006). For each sample, we calculated functional activities of 186 functions using a Normalized Centroid shift method (Yang et al., 2012). For each function j, we classified the 18,071 genes (G) into two classes: genes within the function j (G_{fun_j}) and the other genes (G/G_{fun_j}). We calculated the average rank-based scores $NC_{G_{fun_j}}$ and $NC_{G/G_{fun_j}}$, and then the activity score of function j (FAS_{fun_j}) was calculated as the difference between $NC_{G_{fun_j}}$ and $NC_{G/G_{fun_j}}$.

$$NC_{G_{fun_j}} = \frac{\sum_{g_i \in G_{fun_j}} W_s(g_i)}{\left|G_{fun_j}\right|}$$

$$NC_{G/G_{fun_j}} = \frac{\sum_{g_i \in G \cap g_i \notin G_{fun_j}} W_s(g_i)}{\left|G/G_{fun_j}\right|}$$

$$FAS_{fun_j} = NC_{G_{fun_j}} - NC_{G/G_{fun_j}}$$

Identifying Dysregulated Functions With Significant Activity Difference Between the Engineered Organoids and CRC Samples

To measure the biological distance from engineered organoids (S) and corresponding CRC samples (T), we compared the activities of 186 functions between S and T. For each type of mutation combination, we calculated the average functional activities of each function, $FAS^s_{fun_j}$ and $FAS^T_{fun_j}$, for S and T. The $\mathrm{DFAS}_{fun_j} = \left|FAS^s_{fun_j} - FAS^T_{fun_j}\right|$ measure the activity difference. To determine the significance of activity difference and identify dysregulated functions, the gene expression profiles of S and T were permuted 1,000 times, respectively. We re-calculated 1,000 random DFAS as described above. The significance P was calculated as the frequency in which random DFAS was larger than real DFAS. We identified the dysregulated functions as those at FDR = 0.01.

Inferring Subsequent Key Genes During the Progression of CRC

The known driver mutations were inefficient to capture cancer behaviors and to broadly explain cancer mechanisms. Exploring the subsequent key genes of known driver mutations can improve the understanding of modeling CRC. We utilized Random walk with restart (RWR) (Köhler et al., 2008) to infer subsequent key genes during the progression of CRC for five types of organoids.

For each dysregulated function k obtained from a specific organoid, we reconstructed a biological network based on the pathway structure. We calculated the degrees of genes in the dysregulated function and selected the top 10% genes with the highest degrees as the seed genes which were the input of random

walk. The seed genes were sowed into the protein interaction network. The information flow can restart from the seed genes with probability r in RWR (Köhler et al., 2008):

$$P_{t+1} = (1 - r)WP_t + rP_0$$

where r was set to 0.7; $P_0$was the initial probabilities of genes, in which the probabilities of seed genes was $1/n$ (n was the number of seed nodes) and others 0; $P_t$were the probabilities of genes at the t_{th} steps; W was the normalized transfer matrix of the protein interaction network; the random walk process reached the steady-state when the maximum difference between $P_{t+1}$and $P_t$was $<10^{-8}$. The P_{t+1} characterized the functional similarity of genes with seed genes. We randomly selected 1,000 sets of pseudo seed genes with the same size and re-performed random walk. For each gene j in the protein interaction network, the significance P_j^k was calculated as the frequency in which random functional similarity was larger than real one. Finally, we combined the significance (P_j^k) of gene j calculated from all dysfunctional functions ($k = 1 \ldots\ldots K$) into a statistic X which follow the χ^2 (2K) distribution:

$$\chi^2 = -2\sum_{k=1}^{K} \ln P_j^k$$

Where the K was the number of dysfunctional functions. The $P(X \geq \chi^2 | X \sim \chi^2(2K))$ represented the significance of genes. We considered genes with FDR ≤ 0.05 as subsequent key genes.

Prioritizing Gene Cascading Paths to Recapture the Adenoma-Carcinoma Sequence of CRC

High tumor heterogeneity of genetic alterations in CRC made the well-known adenoma-carcinoma sequence explain a part of CRC patients, additional alternative gene paths were needed to interpret the development progression of more extensive CRC patients. Different patients with similar phenotype had different combinations of genetic alterations that tended to participate in same or similar functions. To prioritize gene cascading paths for each type of organoid, firstly, we calculated the functional coherence among the five known genes and the subsequent key genes (Wang et al., 2007), and constructed the functional consistency network at the threshold of 0.4. Then, a sparse functional consistency network was constructed by selecting two neighbors with highest functional consistency for each gene. Finally, using the well-known adenoma-carcinoma sequence model as the template, each gene cascading path was identified by starting from the mutant genes in the organoids and ending at the potential key gene showing the maximum shortest distance with mutant genes.

Stepwise Comparison of Five Types of Organoids in the Activities of Hallmark Signatures

We compared the activities of 50 hallmark signatures among five types of organoids (including A-organoid, AS-organoid, AK-organoid, AKST-organoid, and AKSTP-organoid) in a stepwise way. For a pair of organoids, we identified the significant activation/inactivation of hallmark signatures in the organoid with more mutations by comparing with the other. The activities of 50 hallmark signatures were estimated using gene set enrichment analysis, and the activity differences between the pair of organoids were calculated. To measure the significance of activity differences, we permutated the transcriptomes of the pair of organoids 1,000 times, and recalculated 1,000 random activity differences of hallmark signatures. The significance of activation was calculated as the frequency in which random activity differences was larger than real one. And the significance of inactivation was calculated as the frequency in which random activity differences was smaller than real one. We identified the significant activation/inactivation of hallmark signatures at FDR ≤ 0.05.

RESULTS

The Combination Mutation Patterns in Five Driver Genes Across CRC Populations

The mutations of five genes (including *APC*, *KRAS*, *SMAD4*, *TP53,* and *PIK3CA*) were reported to play driver roles in CRC progression. Five CRC populations in the cbioPortal were collected to investigate the mutation distributions of the five driver genes (Cerami et al., 2012; Gao et al., 2013). We found that these five genes showed high mutation frequencies ranging from 77 to 100% (**Figure 2A**). As a "gatekeeper" gene, *APC* mutations were extremely pervasive across CRC populations. Especially, the mutation frequency of *APC* reached up to 91% in MSKCC study (**Figure S1** and **Table S2**). The mutation frequencies of *TP53* were 82, 53, 55, 56, and 43% across five CRC populations; 55, 42, 44, 51, and 28% for *KRAS*; 20, 20, 15, 31, and 21% for *SMAD4*; and 12, 14, 15, 24, and 10% for *PIK3CA*. The high frequencies of these five driver genes confirmed their core roles in the progression of CRC. Interestingly, only 0.72, 0.94, 0.45, 0% (0/72), 0% samples harbored the mutations of all five genes across the five CRC populations (**Figure 2B**). CRC samples harboring mutations in four genes only occupied 16.7, 5.7, 5.5, 8.3, and 3.9%, respectively. Most CRC samples (74.6, 65.1, 63.6, 58.3, and 54.1%) carried mutations of two or three genes. And the most common combination of mutations was observed between *APC* and *TP53*. These results further showed CRC was a highly heterogeneous disease from genomic perspective. Different CRC patients harbored different combinations of genetic alterations. The mutation frequency of single driver gene was high while the co-occurrence frequency of the five driver genes was very low. These phenomenon implied that although the mutations of the five driver genes could explain the CRC pathogenesis well, which could only explain the progressive mechanism for a fraction of CRC patients, but the molecular pathogenesis of major patients

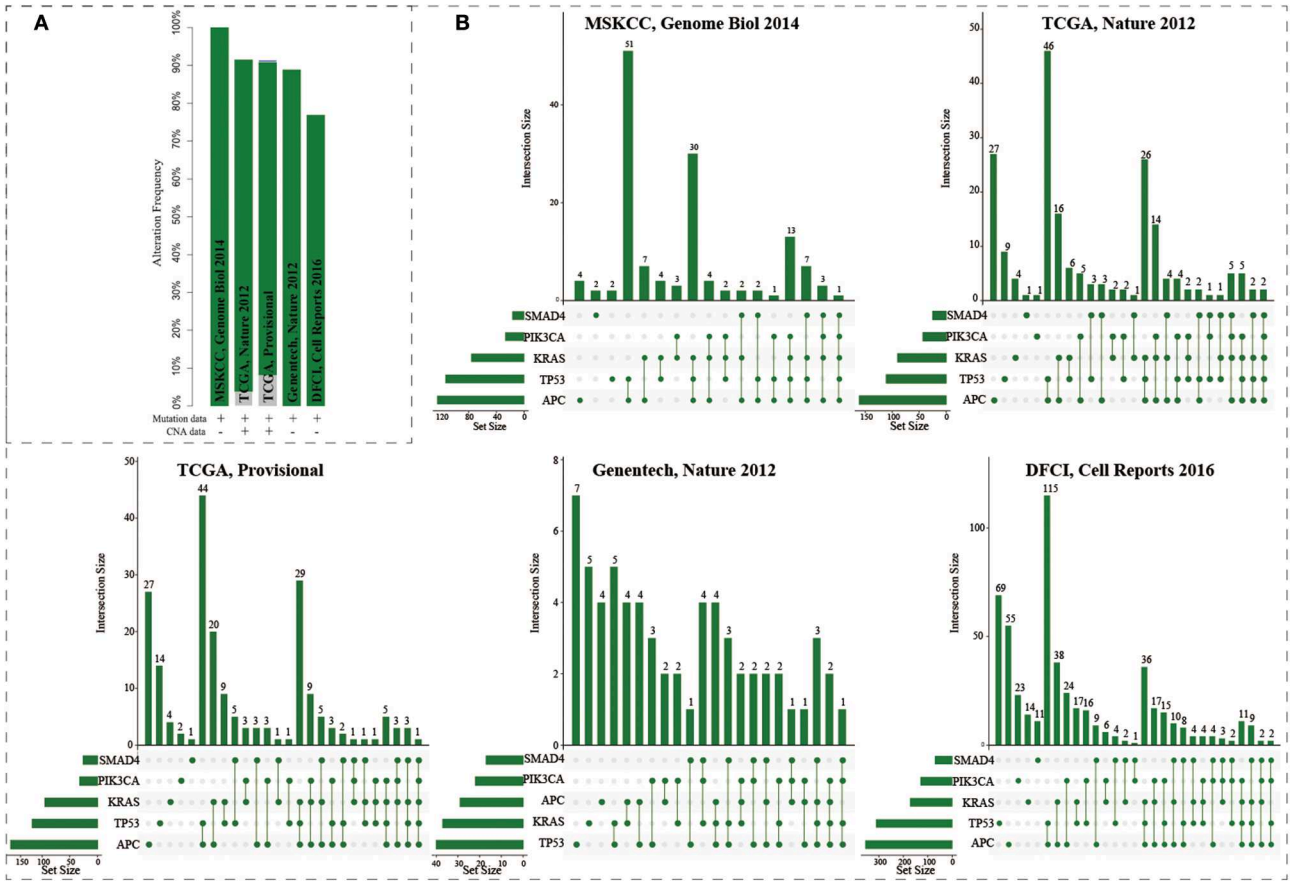


FIGURE 2 | The mutation distributions of five driver genes in the five CRC populations. **(A)** The mutation frequencies of the five genes across five CRC populations. **(B)** Combination patterns of mutations in the five genes across CRC populations.

remains unclear. There existed other gene paths or mutation combinations to drive CRC evolution.

Functionally Characterizing Engineered Organoids Carrying Various Combinations of Driver Mutations

We collected the transcriptomes of five types of engineered organoids which expressed mutations of different combinations of the five genes from GSE57965. For each type of engineered organoid, we calculated the activities of 50 hallmark signatures from MSigDB and identified the hallmark signatures with significant activation or inactivation using gene set enrichment analysis (Subramanian et al., 2005; Liberzon et al., 2015). In A-organoid, epithelial mesenchymal transition was the most significantly activated development signature (**Figure S2A**, $P < 0.001$). The immune signatures [IL6- JAK-STAT3 signaling ($P = 0.0012$) and inflammatory response ($P = 0.001$)] also showed significant activation. Five of six proliferation signatures showed significant activation in AK-organoid, which contained G2M checkpoint ($P < 0.001$) and E2F targets ($P < 0.001$). In AKST- and AKSTP-organoids, the hypoxia and glycolysis signature showed significant activation. Notably, none of 50 hallmark signatures showed significant inactivation in AKSTP-organoid (**Figure S2B**), indicating AKSTP-organoid exhibited more cancer hallmarks. These results suggested that the introduction of the five driver genes in intestinal organoids could induce the activation of hallmark signatures.

Dynamically Analyzing CRC Progression From A- to AKSTP-Organoids

To further characterize the dynamic activities of hallmark signatures during sequential introduction of multiple driver mutations, we compared the activities of hallmark signatures between the five types of organoids. Compared with A-organoids, the other four types of organoids showed consistent activation of proliferation signatures containing G2M checkpoint and E2F targets (**Figure 3A**). Further, compared with AK- and AS-organoids, the AKST- and AKSTP-organoids consistently activated the hypoxia and glycolysis signature (**Figures 3B,C**). Compared with AKST-organoid, the AKSTP-organoid continued to enhance activation of proliferation signatures (MYC targets and P53 pathway) and immune signatures (**Figure 3D**). These dynamic analyses suggested that sequential introduction of these driver mutations gradually drove the activation of distinct

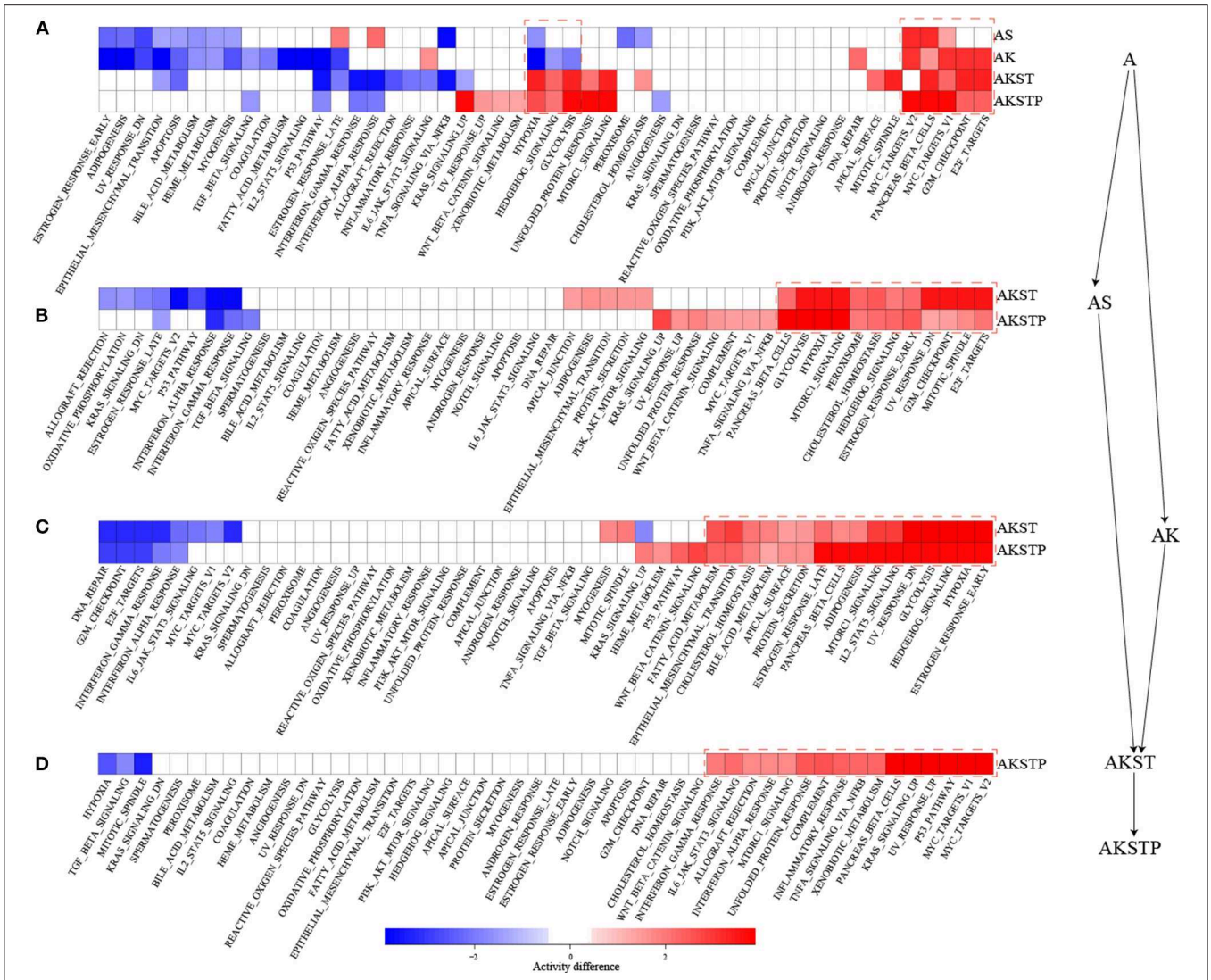


FIGURE 3 | Dynamic activity analysis of 50 hallmark signatures from A-organoid (A) to AKSTP-organoid (AKSTP). **(A)** The significant activation (red) or inactivation (blue) of 50 hallmark signatures in AS-organoid (AS), AK-organoid (AK), AKST-organoid (AKST), and AKSTP-organoid (AKSTP) by comparing with A-organoid (A). **(B)** The significant activation (red) or inactivation (blue) of 50 hallmark signatures in AKST-organoid and AKSTP-organoid by comparing with AS-organoid. **(C)** The significant activation (red) or inactivation (blue) of 50 hallmark signatures in AKST-organoid and AKSTP-organoid by comparing with AK-organoid. **(D)** The significant activation (red) or inactivation (blue) of 50 hallmark signatures in AKSTP-organoid by comparing with AKST-organoid.

hallmark signatures, and conferred the selective advantages to engineered organoids.

Functionally Characterizing Combined Effects of the Five Driver Mutations Using TCGA CRC Patients

We collected CRC samples with both expression and mutation profiles from TCGA. The mutations of the driver genes could influence gene expression levels of driver genes ($P = 0.021$ for *APC*, $P = 0.0174$ for *SMAD4*, $P = 2.7e{-}5$ for *TP53*, $P = 0.0013$ for *KRAS*, and $P = 0.0183$ for *PIK3CA*, **Figure S3**). According to the mutation status of the five driver genes, the 103 CRC samples were grouped into five groups (**Table S3**). To evaluate whether CRC samples with different combinations of driver mutations showed differential activities of hallmark signatures, we calculated the activities of hallmark signatures using single-sample GSEA for each CRC sample (Hänzelmann et al., 2013). For each group, average activities of hallmark signatures were calculated. We found that these five groups showed similar activated patterns (**Figure 4A**). The correlation coefficients of average activities ranged from 0.973 to 0.999 (**Figure 4B**). To further investigate whether the similar activated patterns also exited in all CRC samples, the correlation coefficients among all CRC samples were calculated. We found that all CRC samples still exhibited highly consistent correlation of hallmark signature activities in spite of different combinations of genetic alterations (**Figure 4C**). The results suggested that there existed

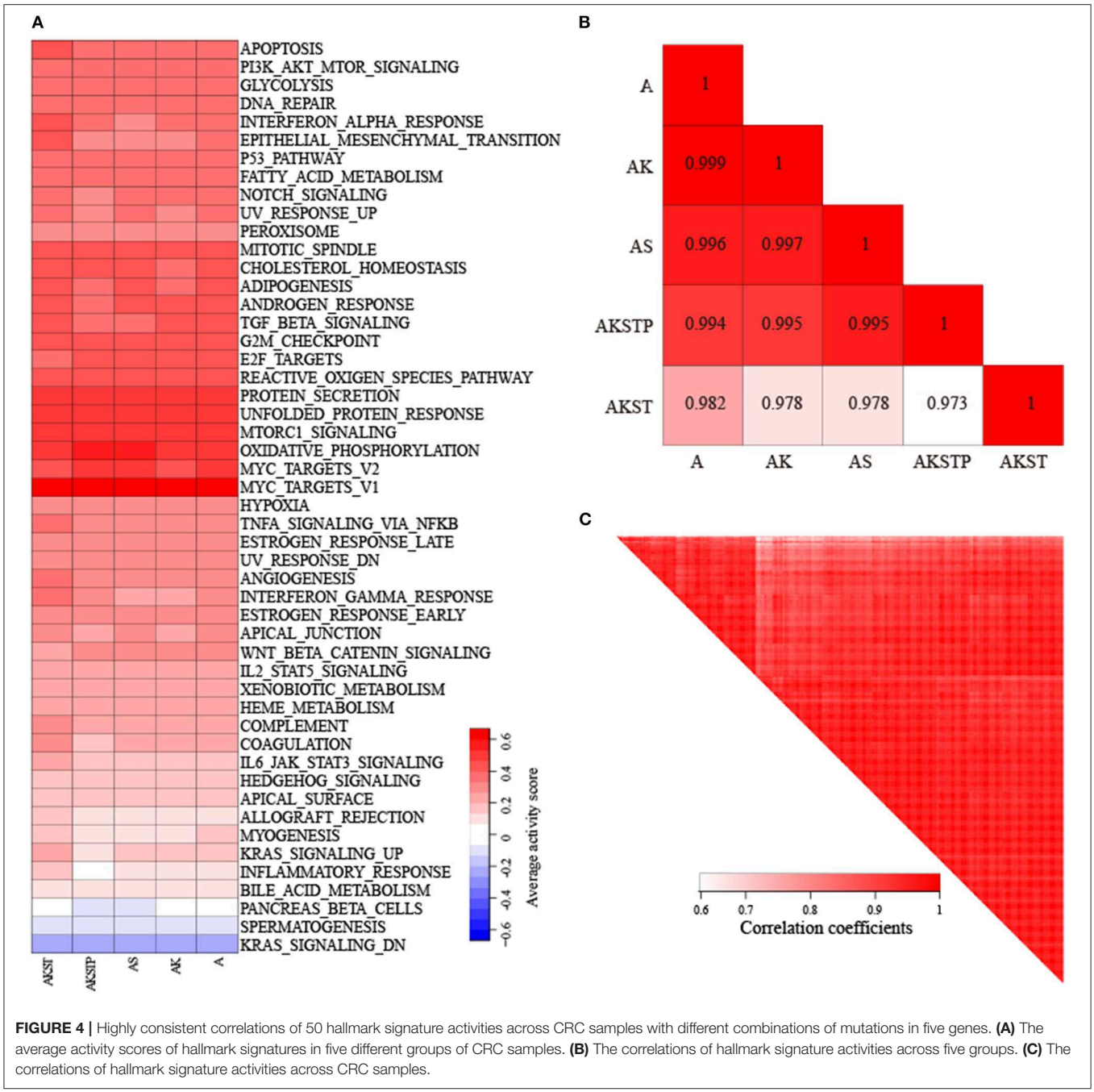


FIGURE 4 | Highly consistent correlations of 50 hallmark signature activities across CRC samples with different combinations of mutations in five genes. **(A)** The average activity scores of hallmark signatures in five different groups of CRC samples. **(B)** The correlations of hallmark signature activities across five groups. **(C)** The correlations of hallmark signature activities across CRC samples.

additional driver genetic alterations contributing to development mechanism of broad CRC patients.

Substantial Biological Gaps Between Engineered Organoids and Colorectal Cancer Tissues

We used the rank-based scores to integrate the expression profiles of engineered organoids and CRC samples. The result of principal components analysis showed that the expression pattern could distinguish the five types of organoids from TCGA CRC samples (**Figure S4**). To characterize the biological distance from the engineered organoids to CRC, we identified the dysregulated functions with significant activity difference between engineered organoids and their corresponding CRC samples at FDR = 0.01 against 1,000 permutations (**Table S4**).

For the A-organoids, we found that 65 of 186 functions showed no significant difference of functional activities by contrast to CRC samples, two of which *APC* participated in directly. For example, *APC* participated in the Wnt signaling pathway directly. In the A-organoids, the WNT pathway showed similar functional activity with the CRC samples with *APC*

mutation ($P = 0.015$, **Figure S5A**). However, the Wnt signaling pathway showed significant activity difference ($P = 0.008$, **Table S4**) by comparing normal and CRC samples. These results suggested that *APC* mutation contributed the activation of Wnt signaling pathway, which was consistent with previous studies (Drost et al., 2015; Matano et al., 2015). Meanwhile, there were 121 dysregulated functions with significant activity difference. The MAPK signaling pathway showed significant activity difference between A-organoids and CRC samples ($P < 0.001$, **Figure S5B**). The number of functions showing similar activities between AK-organoids and corresponding CRC samples were up to 128, and the number of dysregulated functions decreased to 58. The RAS and MAPK signaling pathway showed similar activity between AK-organoids and CRC samples ($P = 0.11$ and $P = 0.33$, **Figures S5C,D**), suggesting the combination of *APC* and *KRAS* mutations enabled the activity of RAS and MAPK signaling pathway to reach the physiological state of CRCs. We also compared the function activity between AS-organoid, AKST-organoids, AKSTP-organoids and their corresponding CRC samples. We found that the number of functions with similar activity increased and the number of dysregulated functions decreased along with the number of genes mutations (**Figure 5A** and **Table S4**). These results gave a clue that combinations of multiple drive mutations approximated the organoids to CRC by activating or inactivating the activities of functions.

To characterize the step-by-step progression of CRCs from organoids engineered by introducing mutations, we compared the activity difference of 186 functions from five types of organoids. Firstly, we focused on the five functions including Wnt signaling pathway, RAS-MAPK signaling pathway, TGF-β signaling pathway, TP53 signaling pathway and PI3K signaling pathway, which were targeted by *APC*, *SMAD4*, *KRAS*, *TP53*, and *PIK3CA*, respectively. By comparing the normal and CRC samples, we found four functions including Wnt, RAS-MAPK, TP53 and PI3K signaling pathway showed significant differential activity($P = 0.008$, $P < 0.001$, $P < 0.001$, and $P = 0.003$, **Table S4**). By introducing the mutations of corresponding genes, we found the significance of activity difference of four functions disappeared gradually (**Table S5**, FDR = 0.01). With the increasing number of mutated genes, the activity difference of these functions between organoids and CRCs tended to random state, suggesting the driver progression of key genes during carcinogenesis.

To further investigate the dynamic progression integrally, we clustered the organoids and the 186 functions based on the significance status of dysregulated functions. We found that A- and AS-organoids were a class, and AK-, AKST-, and AKSTP-organoids as a class (**Figure 5B**). *APC* mutation was a key gene for forming an adenoma. The adenoma still maintained the benign state after introducing *SMAD4* mutation. *KRAS* mutation made the adenoma canceration by dysregulating the activities of many functions, implying *KRAS* mutation played a key role during transformation from adenoma to CRC.

Among the 186 functions, 56 showed no significance of activity difference between any type organoid and CRCs. Twenty one functions also showed no significance between normal and CRC samples, indicating these functions may be essential functions for maintaining cell survival. However, the other 35 functions showed significant activity difference between normal and CRC samples, of which 16 functions were metabolism-related, implying the serious metabolic derangements have occurred from an adenoma. Meanwhile, we found that 27 functions showed significant activity difference between all of five types of organoids and CRCs, such as the PI3K signaling pathway, suggesting that additional key driver mutations were needed to transform the organoids to CRCs.

Prioritizing Gene Cascading Paths Contributing to the Model of Colorectal Cancer Derived From Engineered Organoids

The five driver genes were not sufficient to make organoids approximate the physiological state of CRCs with features of

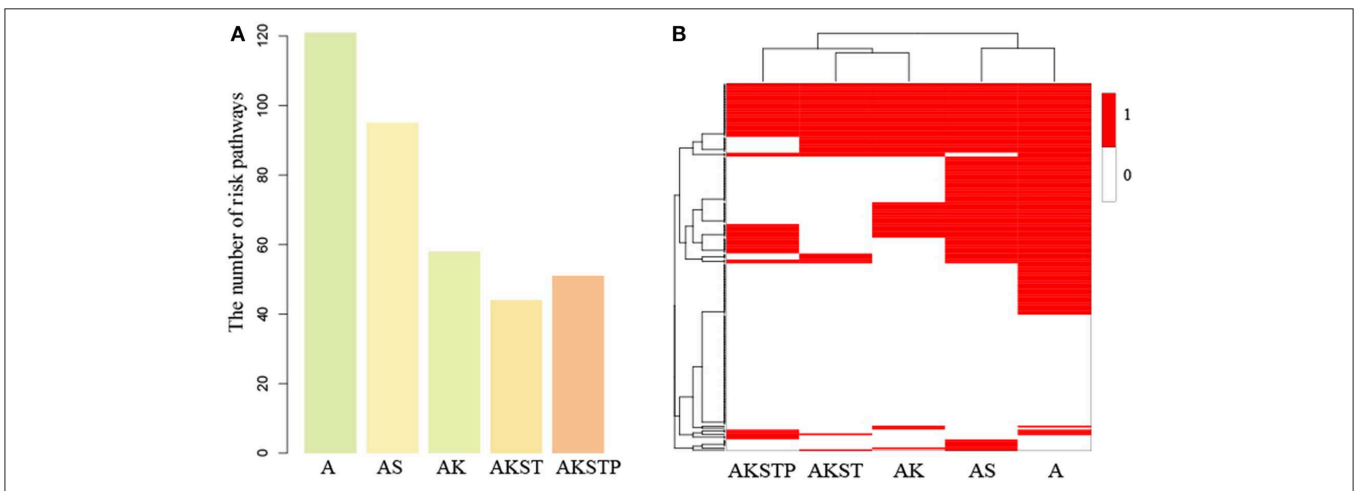


FIGURE 5 | The dysregulated functions identified in the gaps between five types of organoids and CRC samples. **(A)** The number of dysregulated functions identified five types of organoids. **(B)** The binary heatmap of dysregulated functions across five types of organoids. (1 represents dysregulated functions, and 0 represents not).

metastasis and invasion (Matano et al., 2015). Meanwhile, due to tumor heterogeneity of CRCs, the mutations of five driver genes could explain development mechanisms of a part of CRC patients. Additional gene cascading paths were needed to explain the pathogenesis of broad CRC populations.

Using random walk to propagate information flow from dysregulated functions, we identified potential subsequent key genes for five types of organoids. At FDR = 0.05, we predicted 34, 89, 56, 4 potential key genes for A-, AS-, AK-, and AKST-organoids, respectively (**Figures S6A,B** and **Table S6**). For A- and AS-organoids, both *PIK3CA* and *KRAS* were identified, and *PIK3CA* was the top one gene identified from AK- and AKST-organoids, suggesting our method was able to identify key genes (**Figure S6C**). We also found that different organoids needed some common and specific potential genes to compete CRC progression (**Figures S6B,C**).

Heterogeneity in genetic alterations across CRC populations indicated that different combinations of key genes contributed to the tumor progression through participating in similar functions. Prioritizing gene cascading paths for different organoids, which could perform analogical functions of five driver genes, could provide the interpretation of pathogenesis for broader CRC patients. Functional analysis showed the high functional coherence among the five driver genes. We calculated the function coherence among the potential genes and five known genes, and found that many potential key genes showed high functional coherence with the five known genes (**Figures S7–S11**). Thus, using the five driver genes as template, we prioritized cascading paths of key genes based on the function coherence to recapitulate the adenoma-carcinoma sequence model for different organoids (**Figures 6A–E**).

For A-organoids, two paths of potential key genes were predicted: one contained *APC*, *ERBB4*, *NRG1*, *KRAS*, *PIK3CA,* and *PIK3CG*, and the other contained *APC*, *ERBB4*, *LATS2*, *TIAM1*, and *DLC1* (**Figure 6A**). *ERBB4*, one of the ErbB receptor tyrosine kinases, showed the functional coherence of 0.56, 0.59, 0.63, and 0.57 with *APC*, *KRAS*, *PIK3CA,* and *TP53*, respectively, which also participated in cancer associated functions such as MAPK cascade, cell migration and cell proliferation. The colonic inflammation was limited by ErbB4 signaling through stimulating pro-inflammatory macrophage apoptosis (Schumacher et al., 2017). *ERBB4* itself could not induce tumor transformation of mouse colonocytes, while under the condition of colonocytes with mutant Apc and Ras, *ERBB4* enhanced the transformed phenotype both *in vitro* and *in vivo* (Williams et al., 2015). The increased co-expression of ErbB4-CYT-2 with KITENIN promoted the transition of colon adenoma to adenocarcinoma in tumor microenvironment of *APC* loss (Bae et al., 2016). *NRG1*, neuregulin 1, showed the functional coherence of 0.54, 0.58, 0.55, 0.58, and 0.51 with *APC*, *KRAS, PIK3CA*, *SMAD4,* and *TP53*, respectively. In the ERBB signaling pathway, *NRG1* could participate in cell migration and invasion by activating *ERBB4* and *KRAS*, and contribute to cell cycle and cell metabolism by activating *ERBB4* and *PIK3CA*. *NRG1* was methylated in tumors and the knockdown of *NRG1* could increase net cell proliferation (Chua et al., 2009). Paracrine NRG1/HER3 signals promoted CRC cell progression, and was associated with poor prognosis in CRC (De Boeck et al., 2013). *PI3KCG* was a critical switch between immune stimulation and suppression during inflammation and tumor growth (Kaneda et al., 2016). The silencing of *PIK3CG* contributed to inhibit the PI3K-Akt/PKB signaling system which was responsible for the tumorigenesis and progression of colorectal cancers (Semba et al., 2002). Thus, *ERBB4* and *NRG3* may replace *SMAD4* and *TP53* to form a new combination, together with *APC*, *KRAS* and *PIK3CA*, to form an alternative path underlying CRCs.

For ASKT-organoids, *PIK3CA* was ranked first, together with *APC*, *SMAD4*, *KRAS,* and *TP53*, which restored the known the adenoma-carcinoma sequence model of CRC (**Figure 6D**). ASKTP-organoids were capable to form the tumors while showed weak invasive behavior. Additional key genes were needed to complete the progression of CRC. *PKHD1* were the second potential key genes which showed function coherence of 0.47, 0.49, 0.48, 0.45, and 0.45 with *APC, SMAD4*, *KRAS*, *TP53,* and *PIC3CA*, respectively. The protein encoded by *PKHD1* harbored the structural features with hepatocyte growth-factor receptor and plexins which involved in regulation of cell proliferation and cellular adhesion and repulsion (Onuchic et al., 2002). Inhibition of *PKHD1* may control cell cycle via mTOR signaling pathway (Zheng et al., 2009), and induced cell apoptosis through PI3K and NF-κB pathways (Sun et al., 2011). We found that *PKHD1* showed high frequency of mutations in the CRC populations (from 8.9 to 11.8%, **Figure S12**). Previous studies showed that *PKHD1* was a candidate CRC gene by screening mutations in the consensus coding sequences profile, and was assigned to the function of cell adhesion with the first rank (Sjöblom et al., 2006). The germline mutations of *PKHD1* played a protective role in colorectal cancer (Ward et al., 2011). Thus, introduction of *PKHD1* mutations following the five driver genes may contribute to CRC invasion and metastasis.

DISCUSSION

The adenoma-carcinoma sequence was recognized as the mechanism model of CRC, in which mutations of *APC*, *KRAS*, *SMAD4*, *TP53,* and *PIK3CA* could sequentially drive CRC transformation. The sequential introduction of CRC genes was used to model colorectal cancer. These studies gave a clue that it is possible to investigate the CRC dynamic progression using engineered organoids. We proposed an integrative strategy to characterize the dynamic progression of CRC and prioritize gene cascading paths for directing subsequent introductions of key genes.

Dynamic analysis of activities of biological functions showed biological gaps between organoids and CRC tissues. The number of dysregulated functions dropped sharply with the number of mutations of key genes increasing. These results were consistent with previous studies (Drost et al., 2015; Matano et al., 2015), suggesting that our method could capture biological dynamics and characterize the CRC progression. The AKST- and AKSTP- organoids approximated the true CRC with corresponding mutations. However, there were still many dysregulated functions associated

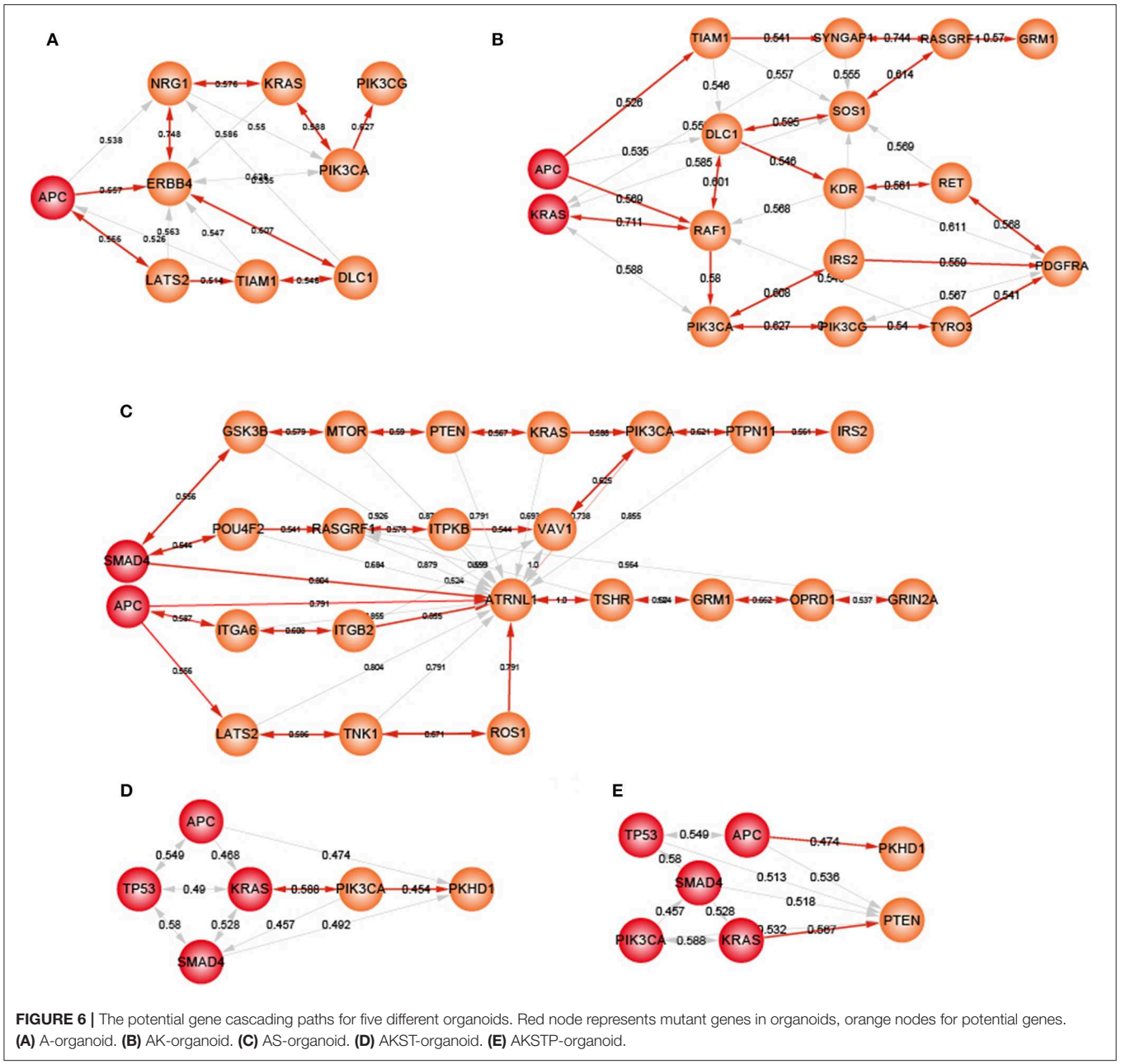


FIGURE 6 | The potential gene cascading paths for five different organoids. Red node represents mutant genes in organoids, orange nodes for potential genes. **(A)** A-organoid. **(B)** AK-organoid. **(C)** AS-organoid. **(D)** AKST-organoid. **(E)** AKSTP-organoid.

with tumor metastasis, such as cytokine-cytokine receptor interaction, ECM-receptor interaction, and adherent junction. Meanwhile, some tumor microenvironment associated functions including antigen processing and presentation, leukocyte transendothelial migration and chemokine signaling pathway were also in these biological gaps. The identified dysregulated functions may provide an explaining that AKST- and AKSTP-organoids without features of migration and invasion may be due to lacking of tumor microenvironment supporting invasion and metastasis. Additional driver mutations of key genes were needed to further identify to control these functions.

Through screening the genetic alteration profiles of CRC populations, the co-occurrence frequency of five CRC genes was low. Although the adenoma-carcinoma sequence of CRC was recognized, it only explained molecular mechanism in a fraction of CRC populations with mutations of all five genes. The genetic alterations of CRC populations showed high heterogeneity, implicating that other key genes were required for drawing the mechanism of colon carcinogenesis for most of CRC populations. Our method not only could characterize biological gaps between different types of organoids and their corresponding CRC samples, but also be able to predict key genes which followed the introduced key mutation to further

shrink biological gaps. The potential sequential genes were identified for different types of organoids, which participated in important functions and pathways. For example, for the AK-organoids, 56 subsequent genes were predicted. Using functional enrichment, many cancer-associated functions, such as MAPK cascade, Ras signaling pathway, PI3K-Akt signaling pathway, positive regulation of cell migration and positive regulation of cell proliferation, were identified (**Table S7**). With the accumulation of published studies about CRC organoids and multidimensional omics data of organoids (Fumagalli et al., 2017; Newey et al., 2019; Ooft et al., 2019), our method could be used to identify more extensive gene paths and construct the landscape of molecular pathogenesis for CRC cancer. Sequential introduction of the mutations in gene paths may provide a new avenue for understanding the dynamic progression of CRC.

In summary, we developed an integrative strategy to capture the dynamic progression of CRC and prioritize gene cascading paths for understanding the mechanisms of wide CRC patients. Our approach also can reveal the dynamic transformation mechanism of other cancer types. This will provide a more detailed interpretation for molecular mechanisms of cancer which could help for drug design and cancer therapy.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE57965(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57965), TCGA(https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

XL and YX designed and guided this work and LW supervised this work. YP, CX, LX, and GL participated in data processing, program implementation, and paper writing. YZ, CD, YL, FY, and JS contributed to data collecting and organized the figures and tables. All authors provided critical advice for the final manuscript.

FUNDING

This work was supported in part by the National High Technology Research and Development Program of China [863 Program, Grant No. 2014AA021102], the National Program on Key Basic Research Project [973 Program, Grant No. 2014CB910504], the National Natural Science Foundation of China [Grant Nos. 61573122, 31601076], the China Postdoctoral Science Foundation (2016M601444), Wu lien-teh youth science fund project of Harbin medical university [Grant No. WLD-QN1407], Special funds for the construction of higher education in Heilongjiang Province [Grant No. UNPYSCT-2018068], the Heilongjiang Postdoctoral Foundation (LBH-Z16119).

ACKNOWLEDGMENTS

The authors acknowledged the contributions of the data used in this work by all the researchers to public database GEO, TCGA, cBioPortal, and MSigDB. All the research results based on the data are the sole responsibility of the authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00012/full#supplementary-material

REFERENCES

Amar, D., Hait, T., Izraeli, S., and Shamir, R., et al. (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Res.* 43, 7779–7789. doi: 10.1093/nar/gkv810

Bae, J. A., Kho, D. H., Sun, E. G., Ko, Y. S., Yoon, S., Lee, K. H., et al. (2016). Elevated coexpression of KITENIN and the ErbB4 CYT-2 Isoform promotes the transition from colon adenoma to carcinoma following APC loss. *Clin. Cancer Res.* 22, 1284–1294. doi: 10.1158/1078-0432.CCR-15-0306

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Brenner, H., Kloor, M., and Pox, C. P. (2014). Colorectal cancer. *Lancet* 383, 1490–1502. doi: 10.1016/S0140-6736(13)61649-9

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chen, H. J., Wei, Z., Sun, J., Bhattacharya, A., Savage, D. J., Serda, R., et al. (2016). A recellularized human colon model identifies cancer driver genes. *Nat. Biotechnol.* 34, 845–851. doi: 10.1038/nbt.3586

Chua, Y. L., Ito, Y., Pole, J. C., Newman, S., Chin, S. F., Stein, R. C., et al. (2009). The NRG1 gene is frequently silenced by methylation in breast cancers and is a strong candidate for the 8p tumour suppressor gene. *Oncogene* 28, 4041–4052. doi: 10.1038/onc.2009.259

Cooks, T., Pateras, I. S., Tarcic, O., Solomon, H., Schetter, A. J., Wilder, S., et al. (2013). Mutant p53 prolongs NF-κB activation and promotes chronic inflammation and inflammation-associated colorectal cancer. *Cancer Cell* 23, 634–646. doi: 10.1016/j.ccr.2013.03.022

De Boeck, A., Pauwels, P., Hensen, K., Rummens, J. L., Westbroek, W., Hendrix, A., et al. (2013). Bone marrow-derived mesenchymal stem cells promote colorectal cancer progression through paracrine neuregulin 1/HER3 signalling. *Gut* 62, 550–560. doi: 10.1136/gutjnl-2011-301393

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Drost, J., van Jaarsveld, R. H., Ponsioen, B., Zimberlin, C., van Boxtel, R., Buijs, A., et al. (2015). Sequential cancer mutations in cultured human intestinal stem cells. *Nature* 521, 43–47. doi: 10.1038/nature14415

Fearon, E. R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* 61, 759–767. doi: 10.1016/0092-8674(90)90186-i

Fumagalli, A., Drost, J., Suijkerbuijk, S. J., van Boxtel, R., de Ligt, J., Offerhaus, G. J., et al. (2017). Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. *Proc. Natl. Acad. Sci. U.S.A.* 114, E2357–E2364. doi: 10.1073/pnas.1701219114

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:pl1. doi: 10.1126/scisignal.2004088

Gatza, M. L., Lucas, J. E., Barry, W. T., Kim, J. W., Wang, Q., Crawford, M. D., et al. (2010). A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6994–6999. doi: 10.1073/pnas.0912708107

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Hao, Y., Samuels, Y., Li, Q., Krokowski, D., Guan, B. J., Wang, C., et al. (2016). Oncogenic PIK3CA mutations reprogram glutamine metabolism in colorectal cancer. *Nat. Commun.* 7:11971. doi: 10.1038/ncomms11971

Heckl, D., Kowalczyk, M. S., Yudovich, D., Belizaire, R., Puram, R. V., McConkey, M. E., et al. (2014). Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. *Nat. Biotechnol.* 32, 941–946. doi: 10.1038/nbt.2951

Kaneda, M. M., Messer, K. S., Ralainirina, N., Li, H., Leem, C. J., Gorjestani, S., et al. (2016). PI3Kγ is a molecular switch that controls immune suppression. *Nature* 539, 437–442. doi: 10.1038/nature19834

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database—2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* 4:e1000217. doi: 10.1371/journal.pcbi.1000217

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., et al. (2009). SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.* 37:e131. doi: 10.1093/nar/gkp667

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., et al. (2015). Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat. Med.* 21, 256–262. doi: 10.1038/nm.3802

Morizane, R., Lam, A. Q., Freedman, B. S., Kishi, S., Valerius, M. T., and Bonventre, J. V. (2015). Nephron organoids derived from human pluripotent stem cells model kidney development and injury. *Nat. Biotechnol.* 33, 1193–1200. doi: 10.1038/nbt.3392

Nakayama, M., Sakai, E., Echizen, K., Yamada, Y., Oshima, H., Han, T. S., et al. (2017). Intestinal cancer progression by mutant p53 through the acquisition of invasiveness associated with complex glandular formation. *Oncogene* 36, 5885–5896. doi: 10.1038/onc.2017.194

Newey, A., Griffiths, B., Michaux, J., Pak, H. S., Stevenson, B. J., Woolston, A., et al. (2019). Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. *J Immunother. Cancer* 7:309. doi: 10.1186/s40425-019-0769-8

Onuchic, L. F., Furu, L., Nagasawa, Y., Hou, X., Eggermann, T., Ren, Z., et al. (2002). PKHD1, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple immunoglobulin-like plexin-transcription-factor domains and parallel beta-helix 1 repeats. *Am. J. Hum. Genet.* 70, 1305–1317. doi: 10.1086/340448

Onuma, K., Ochiai, M., Orihashi, K., Takahashi, M., Imai, T., Nakagama, H., et al. (2013). Genetic reconstitution of tumorigenesis in primary intestinal cells. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11127–11132. doi: 10.1073/pnas.1221926110

Ooft, S. N., Weeber, F., Dijkstra, K. K., McLean, C. M., Kaing, S., van Werkhoven, E., et al. (2019). Patient-derived organoids can predict response to chemotherapy in metastatic colorectal cancer patients. *Sci. Transl. Med.* 11:eaay2574. doi: 10.1126/scitranslmed.aay2574

O'Rourke, K. P., Loizou, E., Livshits, G., Schatoff, E. M., Baslan, T., Manchado, E., et al. (2017). Transplantation of engineered organoids enables rapid generation of metastatic mouse models of colorectal cancer. *Nat. Biotechnol.* 35, 577–582. doi: 10.1038/nbt.3837

Platt, R. J., Chen, S., Zhou, Y., Yim, M. J., Swiech, L., Kempton, H. R., et al. (2014). CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* 159, 440–455. doi: 10.1016/j.cell.2014.09.014

Powell, S. M., Zilz, N., Beazer-Barclay, Y., Bryan, T. M., Hamilton, S. R., Thibodeau, S. N., et al. (1992). APC mutations occur early during colorectal tumorigenesis. *Nature* 359, 235–237. doi: 10.1038/359235a0

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* 8, 2281–2308. doi: 10.1038/nprot.2013.143

Riemer, P., Rydenfelt, M., Marks, M., van Eunen, K., Thedieck, K., Herrmann, B. G., et al. (2017). Oncogenic β-catenin and PIK3CA instruct network states and cancer phenotypes in intestinal organoids. *J. Cell Biol.* 216, 1567–1577. doi: 10.1083/jcb.201610058

Roper, J., Tammela, T., Cetinbas, N. M., Akkad, A., Roghanian, A., Rickelt, S., et al. (2017). *In vivo* genome editing and organoid transplantation models of colorectal cancer and metastasis. *Nat. Biotechnol.* 35, 569–576. doi: 10.1038/nbt.3836

Rosenfeldt, M. T., O'Prey, J., Morton, J. P., Nixon, C., MacKay, G., Mrowinska, A., et al. (2013). p53 status determines the role of autophagy in pancreatic tumour development. *Nature* 504, 296–300. doi: 10.1038/nature12865

Sánchez-Rivera, F. J., Papagiannakopoulos, T., Romero, R., Tammela, T., Bauer, M. R., Bhutkar, A., et al. (2014). Rapid modelling of cooperating genetic events in cancer through somatic genome editing. *Nature* 516, 428–431. doi: 10.1038/nature13906

Schumacher, M. A., Hedl, M., Abraham, C., Bernard, J. K., Lozano, P. R., Hsieh, J. J., et al. (2017). ErbB4 signaling stimulates pro-inflammatory macrophage apoptosis and limits colonic inflammation. *Cell Death Dis.* 8:e2622. doi: 10.1038/cddis.2017.42

Semba, S., Itoh, N., Ito, M., Youssef, E. M., Harada, M., Moriya, T., et al. (2002). Down-regulation of PIK3CG, a catalytic subunit of phosphatidylinositol 3-OH kinase, by CpG hypermethylation in human colorectal carcinoma. *Clin. Cancer Res.* 8, 3824–3831.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268–274. doi: 10.1126/science.1133427

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, L., Wang, S., Hu, C., and Zhang, X. (2011). Down-regulation of PKHD1 induces cell apoptosis through PI3K and NF-κB pathways. *Exp. Cell Res.* 317, 932–940. doi: 10.1016/j.yexcr.2011.01.025

Takeda, H., Kataoka, S., Nakayama, M., Ali, M. A. E., Oshima, H., Yamamoto, D., et al. (2019). CRISPR-Cas9-mediated gene knockout in intestinal tumor organoids provides functional validation for colorectal cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15635–15644. doi: 10.1073/pnas.1904714116

Tang, J., Feng, Y., Kuick, R., Green, M., Green, M., Sakamoto, N., et al. (2019). Trp53 null and R270H mutant alleles have comparable effects in regulating invasion, metastasis, and gene expression in mouse colon tumorigenesis. *Lab. Invest.* 99, 1454–1469. doi: 10.1038/s41374-019-0269-y

van de Wetering, M., Francies, H. E., Francis, J. M., Bounova, G., Iorio, F., Pronk, A., et al. (2015). Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* 161, 933–945. doi: 10.1016/j.cell.2015.03.053

van Lidth de Jeude, J. F., Meijer, B. J., Wielenga, M. C. B., Spaan, C. N., Baan, B., Rosekrans, S. L., et al. (2017). Induction of endoplasmic reticulum stress by deletion of Grp78 depletes Apc mutant intestinal epithelial stem cells. *Oncogene* 36, 3397–3405. doi: 10.1038/onc.2016.326

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799. doi: 10.1038/nm1087

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Ward, C. J., Wu, Y., Johnson, R. A., Woollard, J. R., Bergstralh, E. J., Cicek, M. S., et al. (2011). Germline PKHD1 mutations are protective against colorectal cancer. *Hum. Genet.* 129, 345–349. doi: 10.1007/s00439-011-0950-8

Williams, C. S., Bernard, J. K., Demory Beckler, M., Almohazey, D., Washington, M. K., Smith, J. J., et al. (2015). ERBB4 is over-expressed in human colon

cancer and enhances cellular transformation. *Carcinogenesis* 36, 710–718. doi: 10.1093/carcin/bgv049

Xue, W., Chen, S., Yin, H., Tammela, T., Papagiannakopoulos, T., Joshi, N. S., et al. (2014). CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature* 514, 380–384. doi: 10.1038/nature13589

Yang, X., Regan, K., Huang, Y., Zhang, Q., Li, J., Seiwert, T. Y., et al. (2012). Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput. Biol.* 8:e1002350. doi: 10.1371/journal.pcbi.1002350

Zheng, R., Wang, L., Fan, J., and Zhou, Q. (2009). Inhibition of PKHD1 may cause S-phase entry via mTOR signaling pathway. *Cell Biol. Int.* 33, 926–933. doi: 10.1016/j.cellbi.2009.06.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

REVIEW
published: 05 February 2020
doi: 10.3389/fonc.2020.00068

Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis

*Hong Zheng[1†], Guosen Zhang[1†], Lu Zhang[1], Qiang Wang[1], Huimin Li[1], Yali Han[1], Longxiang Xie[1], Zhongyi Yan[1], Yongqiang Li[1], Yang An[1], Huan Dong[1], Wan Zhu[2] and Xiangqian Guo[1]**

[1] Cell Signal Transduction Laboratory, Bioinformatics Center, School of Basic Medical Sciences, School of Software, Institute of Biomedical Informatics, Henan University, Kaifeng, China, [2] Department of Anesthesia, Stanford University, Stanford, CA, United States

Prognostic biomarkers are of great significance to predict the outcome of patients with cancer, to guide the clinical treatments, to elucidate tumorigenesis mechanisms, and offer the opportunity of identifying therapeutic targets. To screen and develop prognostic biomarkers, high throughput profiling methods including gene microarray and next-generation sequencing have been widely applied and shown great success. However, due to the lack of independent validation, only very few prognostic biomarkers have been applied for clinical practice. In order to cross-validate the reliability of potential prognostic biomarkers, some groups have collected the omics datasets (i.e., epigenetics/transcriptome/proteome) with relative follow-up data (such as OS/DSS/PFS) of clinical samples from different cohorts, and developed the easy-to-use online bioinformatics tools and web servers to assist the biomarker screening and validation. These tools and web servers provide great convenience for the development of prognostic biomarkers, for the study of molecular mechanisms of tumorigenesis and progression, and even for the discovery of important therapeutic targets. Aim to help researchers to get a quick learning and understand the function of these tools, the current review delves into the introduction of the usage, characteristics and algorithms of tools, and web servers, such as LOGpc, KM plotter, GEPIA, TCPA, OncoLnc, PrognoScan, MethSurv, SurvExpress, UALCAN, etc., and further help researchers to select more suitable tools for their own research. In addition, all the tools introduced in this review can be reached at http://bioinfo.henu.edu.cn/WebServiceList.html.

Keywords: web server, tool, prognosis, survival, cancer

OPEN ACCESS

Edited by:
Pasquale Simeone, Università degli Studi G. d'Annunzio Chieti e Pescara, Italy

Reviewed by:
Daniele Vergara, University of Salento, Italy
Chenkai Ma, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

***Correspondence:**
Xiangqian Guo
xqguo@henu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *31 October 2019*
Accepted: *15 January 2020*
Published: *05 February 2020*

Citation:
Zheng H, Zhang G, Zhang L, Wang Q, Li H, Han Y, Xie L, Yan Z, Li Y, An Y, Dong H, Zhu W and Guo X (2020) Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis. Front. Oncol. 10:68. doi: 10.3389/fonc.2020.00068

INTRODUCTION

The prognosis estimation of tumor patient is of great significance to guide clinical treatments and facilitate the elucidation of tumorigenesis mechanism. In current clinical practice, prognosis is determined by many factors, such as disease stage, clinical performance, treatment experience and understanding of the cancer development. However, these properties are relative subjective and may lead to inaccurate prognostic estimates, and may even lead to inappropriate anticancer management strategy. Genotype-Tissue Expression (GTEx) and the Cancer Genome Atlas (TCGA) projects offer a large number of RNA sequence data of normal and cancer samples, providing

unprecedented opportunities for many fields such as cancer bioinformatics and precision medicine to improve our understanding in cancer development and treatment (1, 2). Molecular prognostic biomarkers are the basic components of precision medicine. Data mining and other biological analysis make it possible to predict the prognosis of tumors at the molecular level (3–5). Accurate clinical estimation using prognostic biomarkers helps determining optimal anti-cancer treatment. At the same time, it provides assistance in developing more detailed hospice care plans. So in recent years, the discovery of prognostic biomarkers has become a hot topic in precision medicine.

Numerous studies have evidenced that molecular markers in DNA, RNA and protein level can be as prognostic biomarkers in cancer, and guide the effect of treatment either independently or in addition with present prognosis systems (6–8). In these study, Kaplan-Meier method and multivariate Cox proportional hazards regression models were commonly used to evaluate the associations between molecular markers and survival of patients with cancer (9, 10). However, these biomarkers are not suitable for clinical application due to the lack of independent validation and poor repeatability between different studies.

Mining data from public datasets and making assessments and predictions can be challenging and time-consuming. To extract useful information from these datasets, it requires researchers with strong bioinformatics expertise. To allow more researchers be able to quickly extract information they need, online tools that can easily perform survival analysis from these data are needed. The rapid growth of public datasets has enabled some research groups to focus on collecting omics datasets and developing online bioinformatics prognostic tools and web servers. These various prognostic analysis tools provide valuable evidence and ideas for cancer researchers. However, for many researchers and clinicians, it may be difficult to find the most suitable tool for their own research quickly. This review attempts to provide a comprehensive overview of the commonly used online prognostic tools for cancer prognostic analysis. In addition, the main challenges and future directions in this field are also discussed in this paper.

MATERIALS AND METHODS

Literature research and data collection: the survival analysis tools reviewed in this paper include online prognostic bioinformatics tools and web servers developed by applying different types of profiling data (genomics, epigenomics, proteomics etc.) from clinical samples of different cohorts. Search Strategy for prognostic tools was executed in PubMed and Google Scholar from Jan 1, 2000 to August 31, 2019. Search terms include: "survival analysis," "web server," "prognostic biomarker" and "cancer," keywords combination was used for search. The search was limited to English language. There are 886 articles that matched to above criteria. In the review, 22 representative databases that can be used for the prognosis analysis of multiple cancer types were selected for detailed description; because most of the prognostic tools for single type of cancer were included in the above databases, so we just gave a brief introduction. Ten of these databases are based on mRNA profiling data for prognostic analysis, three databases based on ncRNA profiling data, two databases based on protein data, two databases based on DNA data, and five databases based on multi-omics data. The literature retrieval process is shown in **Figure 1**. The release time of prognostic databases is presented in **Figure 2**. The date of the last search and collating data for these databases was December 10, 2019.

RESULTS

Web Servers for Survival Analysis Based on mRNA Data

In the past two decades, high-throughput gene chips and next-generation sequencing technologies have provided opportunities to explore important cancer-related molecules, therapeutic targets, diagnostic, and prognostic biomarkers. With the implementation of the Cancer Genome Atlas (TCGA) project, a large number of epigenome, transcriptome, and proteome data of tumor samples became publicly accessible. Researchers can analyze the correlation between these data and survival, and look for prognostic biomarkers. Many studies have shown that mRNA expression is closely related to cancer prognosis (11–13). In order to promote the development and evaluation of prognostic biomarkers, some research groups have developed prognosis tools and web servers based on mRNA data by mining TCGA and GEO (Gene Expression Omnibus) data and adding complex statistical calculation. This review introduces 14 bioinformatics tools for evaluating cancer prognosis based on mRNA data (**Table 1**).

LOGpc[1]

LOGpc is a web server that contains a large number of datasets for survival analysis, which provides 13 types of survival terms for 28,098 cancer patients from 26 types of malignant tumors, including OSlms, OSblca, OSkirc and other 23 online prognostic tools (14–21). These patient samples were collected mainly from TCGA and GEO cohorts. LOGpc is free and easy to operate. Twenty six types of tumors are classified into 11 system categories according to TCGA. Currently, only official gene symbol input is acceptable in LOGpc. When user input the gene symbol and set the relative parameters, then click on the "Kaplan-Meier plot" button and the results will be displayed on the output webpage. In order to meet the specific needs from different researchers, clinical confounding factors can also be defined for advanced subgroup analysis.

GENT2[2]

GENT2 provides the differential expression analysis and prognosis analysis based on tumor subtypes (22). The users can search the gene expression profiles of different tissues, and compare the expression levels between tissue subtypes. For survival analysis, this tool provides Kaplan Meier plot with log

[1]http://bioinfo.henu.edu.cn/DatabaseList.jsp
[2]http://gent2.appex.kr

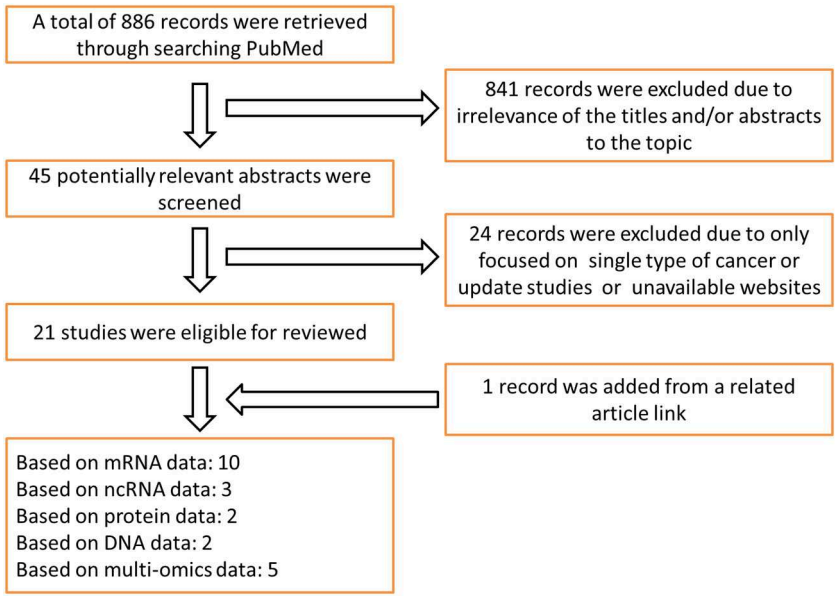


FIGURE 1 | Search flowchart: prognostic web servers for cancers included and excluded in each step.

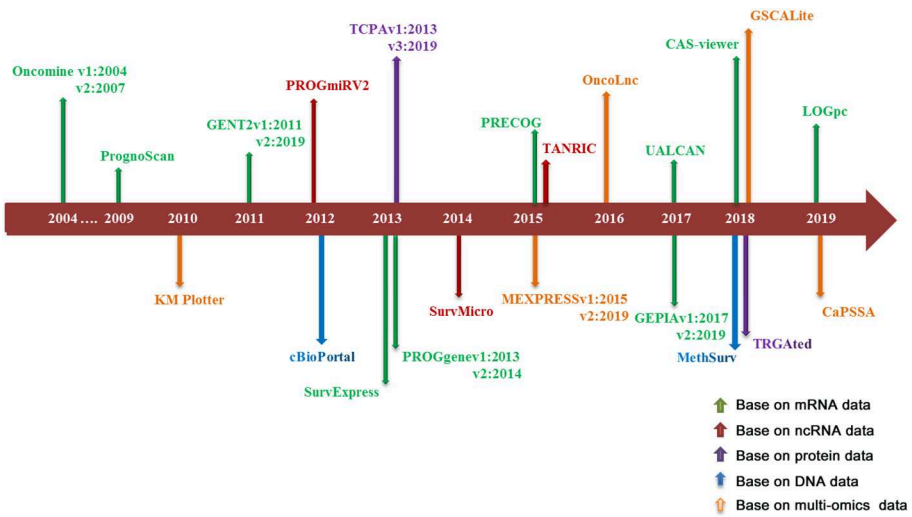


FIGURE 2 | The time axis for the publication of prognostic web servers.

rank test and establishing Cox proportional risk model for meta-analysis. At present, it provides survival analysis for 27 cancer types, including 46 subtypes of 19 cancer types.

PROGgeneV2[3]

PROGgeneV2 is a web-based tool for studying the prognosis of genes in a variety of cancers (23, 24). In current it comprises 193 datasets for 27 cancer types. The users can perform survival analysis of single gene, multi genes and two genes expression ratio, and also use the function of adjusting covariate survival model. Users can upload customized gene datasets for survival analysis of interested genes and compare the results with previously published studies.

SurvExpress[4]

SurvExpress is for studying risk assessment and survival analysis. It contains more than 29,000 samples of 26 cancer types with clinical information from 144 datasets (25). The outputs generated by SurvExpress include the Kaplan-Meier plots by risk group, a heat map of gene expression values and a visual

[3]http://genomics.jefferson.edu/proggene/

[4]http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp

TABLE 1 | Comparison of prognostic web servers based on mRNA data.

Web server	Datasets	Cancer types	Samples	Subgroup analysis	Multi-gene query	Optimal cut-off	Login required
LOGpc	193	26	28,098	Yes	No	No	No
GENT2	195	27	–	Yes	No	No	No
PROGgeneV2	193	27	28,503	Yes	Yes	No	No
SurvExpress	144	26	29,110	Yes	Yes	No	No
PRECOG	165	39	19,168	Yes	No	No	Yes
Oncomine	103	25	17,217	Yes	No	No	Yes
PrognoScan	74	23	9,196	No	No	Yes	No
KM Plotter	45	21	12,984	Yes	Yes	Yes	No
GSCALite	63	33	10,558	Yes	Yes	No	No
UALCAN	35	31	7,233	Yes	Yes	No	No
GEPIA	33	33	10,558	No	Yes	No	No
CAS-viewer	33	33	10,558	Yes	No	No	No
MEXPRESS	33	33	–	Yes	No	No	No
CaPSSA	28	27	10,206	No	Yes	No	No
OncoLnc	21	21	8,616	No	No	No	No

–, survival sample data is not displayed on the website.

association of available clinical information to risk groups. Survival ROC estimates the specificity and time-dependent sensitivity for survival risk groups.

PRECOG[5]

PRECOG is a system for integrating genomic profiles and cancer clinical data, it covers 39 different cancer types, including about 19,000 samples with overall survival data from 165 cancer expression datasets (26). It allows researchers to query whether gene expression correlates with patient survival. For simple display, 39 different histologic types of tumors were divided into 18 groups. The correlation between gene expression and overall survival was assessed by univariate Cox regression. PRECOG also provides gene prognosis analysis for pan-cancer. However, new users need to register and log in.

Oncomine[6]

Oncomine is a cancer gene chip database and integrated data mining platform, aiming at mining cancer gene information (27, 28). Oncomine has more complete cancer mutation spectrum, gene expression data and related clinical information, which provides insights to identify new biomarkers or new therapeutic targets. With Oncomine, users can get the results of differential expression, co-expression analysis, molecular concepts analysis, interaction network, correlation analysis between gene expression and survival status, but Kaplan-Meier plot isn't displayed directly. Meta-analysis can also be used to compare various studies to determine more reliable and consistent results. Oncomine Research Edition is free, but needs a valid academic email address to register and log in.

PrognoScan[7]

Prognoscan is a platform for predicting the relationship between gene expression and patient survival based on a large number of public cancer microarray datasets with clinical information. It provides a variety of survival terms for 14 cancer types (29). One of its advantages is that survival analysis in this tool performs the minimum *P*-value method and optimal cut-off is provided.

KMplotter[8]

The Kaplan Meier plotter (KMplotter) can be used for single gene or multiple gene prognosis analysis for many kinds of malignant tumors (30–32). Researchers can assess the effect of mRNA and miRNA expression on the survival rate of 21 cancer types by pan-cancer analysis. When the users input the relevant gene name and select the appropriate gene expression cut-off point, the comparison results between the two groups will be displayed with 95% confidence interval, risk ratio and log rank *P*-value. An Auto best cut-off is provided to compute all possible cut-off values to get the best performing threshold in survival analysis.

GSCALite[9]

GSCALite is a tool for analyzing expression/variation/ clinical correlation of gene sets in cancers with dynamic and visualization manner (33). It provides three survival analysis modules for a gene set based on cancer multi-omics data of TCGA. (1) Differential mRNA expression of gene set between tumor and matched normal samples, gene expression between subtypes of each selected cancer, and its effect on overall survival rate. (2) The influence of SNV (single nucleotide variants) frequency and mutation type of gene set on the overall survival rate in a cancer type. (3) Differential expression of methylation between tumor and matched normal samples, and the effect on the survival rate

[5]https://precog.stanford.edu/
[6]http://www.oncomine.org/
[7]http://www.prognoscan.org/
[8]http://kmplot.com/analysis/
[9]http://bioinfo.life.hust.edu.cn/web/GSCALite/

of selected cancer types. It allows users to search for prognostic markers at transcriptome level, epigenetic modification, and DNA mutation. Users can query the cancer pathway activity related to gene expression and the correlation between genes and drug sensitivity, it is convenient for researchers to study drug resistance of tumor.

UALCAN[10]

UALCAN is a web-based tool for analyzing TCGA RNA-seq and clinical data to evaluate the association of gene expression and patient survival, allows users to conduct differential expression analysis and survival analysis for interested genes and access the expression and survival information of a given gene in 31 types of cancers by performing pan-cancer analysis (34). Currently, UALCAN provides protein differential expression analysis for breast cancer, colon cancer, and other three cancer types, but does not provide survival analysis based on protein data. UALCAN also provides additional information about the selected genes or targets by linking to Pubmed, TargetScan, DRUGBANK, and so on, this helps researchers collect more valuable information and data.

GEPIA[11]

GEPIA is an interactive web-based tool for survival analysis based on gene expression, it offer the choice of selecting overall survival (OS) or disease-free survival (DFS) for the analysis (35, 36). According to the characteristics of gene normalization, GEPIA allows two different genes to be input at the same time for survival analysis. GEPIA also presents the top genes most related to the survival of cancer patients. This function is very helpful for the users. In addition to providing patient survival analysis, GEPIA has other functions such as differential expression analysis based between different cancer types, multiple gene comparison, similar genes detection.

CAS-Viewer[12]

CAS-viewer is a web-based tool for multiple level comprehensive analysis by integrating multi-omics data such as mRNA, miRNA, methylation, SNP, and clinical information across different cancer types (37). It links the differential transcriptional expression rate with methylation, miRNA, and splicing regulatory elements of 33 cancer types. "Clinical correlation" module presents Kaplan Meier plot showing the correlation between PSI (percent spliced in) value and survival rate, and in this way users can identify potential transcripts related to different survival outcomes of each cancer type.

MEXPRESS[13]

MEXPRESS is an intuitive web tool for analysis of gene expression, DNA methylation, and association with clinical information including patient survival (38). It provides a very different visual interface, allows users to compare specific genomic features (such as DNA methylation) with gene expression and clinical information. Researchers can study the relationship between DNA methylation and gene expression and multiple clinical variables by using MEXPRESS platform.

CaPSSA[14]

CaPSSA supports users to detect the prognostic value of patient subgroups based on gene expression, mutation or genomic alterations of query genes (39). Importantly, it also supports custom histochemical data analysis with clinical information. For candidate gene sets that user-supplied, interactive patient stratification is supported based on gene expression profiles and genomic alterations, the results of log-rank test and Kaplan Meier plots will be displayed for evaluating the prognostic value.

Web Servers for Studying Prognostic Implications of ncRNA

In the past decade, a large number of studies have shown that non-coding RNA (ncRNA) plays an increasingly important role in epigenetic regulation. ncRNAs involved in the network can affect many molecular targets which are related to the development of cancer, and many ncRNAs are considered as driving factors or suppressors of carcinogenesis (40). MicroRNA (miRNA) as one type of ncRNAs regulates mRNA at the transcriptional or post-transcriptional level (41). Studies have shown that lncRNA (long non-coding RNA) plays an important role in many life activities such as dose compensation effect, epigenetic regulation, cell cycle and cell differentiation, and has become a hot spot in tumor genetics research (42). Their expression in cancer has been studied by high-throughput methods, generating valuable sources of public available datasets. An important step in developing ncRNA biomarkers is to evaluate them in independent cohorts. To help and simplify the assessment of ncRNA signatures in cancer prognosis, several ncRNA prognostic databases have been developed by some research teams using public profiling data (**Table 2**).

PROGmiRV2[15]

PROGmiRV2 is a pan-cancer miRNA prognostics database, whose miRNA data comes from GEO and TCGA (43). Compared with version 1, the datasets and samples of the new version have increased greatly, prognosis analysis has been improved from single cancer type analysis to pan-cancer analysis, and the survival indicators provided have increased from one to three (overall survival, recurrence free survival, and metastasis free survival). Users are also allowed to upload their own customized dataset for prognosis analysis, but registration and login are required.

SurvMicro[16]

SurvMicro is a bioinformatics tool for analyzing cancer prognosis based on miRNA. Its data comes from GEO, TCGA, and ArrayExpress (44). SurvMicro comprises 43 datasets and more than 6,000 samples in 15 different cancer types. Cox multiple fitting was used to evaluate the risk of prognosis, the prognosis

[10]http://ualcan.path.uab.edu/index.html
[11]http://gepia.cancer-pku.cn/
[12]http://genomics.chpc.utah.edu/cas/
[13]https://mexpress.be
[14]http://capssa.ewha.ac.kr
[15]http://xvm145.jefferson.edu/progmir/
[16]http://bioinformatica.mty.itesm.mx/SurvMicro

TABLE 2 | Summary of prognostic web servers based on ncRNA data.

Web server	Datasets	Cancer types	Samples	Subgroup analysis	Biomarker	Multi-gene query	Optimal cut-off	Login required
PROGmiRV2	134	33	19,025	Yes	miRNA	Yes	No	No
SurvMicro	43	15	6,412	Yes	miRNA	No	No	No
KM Plotter	25	21	10,613	Yes	miRNA	Yes	Yes	No
OncoLnc	21	21	8,648	No	miRNA	No	No	No
TANRIC	23	20	6,763	Yes	LncRNA	–	–	No
OncoLnc	18	18	8,023	No	LncRNA	No	No	No

–, related information is not displayed on the website.

TABLE 3 | Comparison of prognostic web servers based on protein data.

Web server	Datasets	Cancer types	Samples	Proteins	Subgroups	Multi-gene query	Optimal cut-off	Login required
TCPAv3.0	35	33	8,328	258	No	No	No	No
TRGAted	31	31	7,843	245	Yes	Yes	Yes	No

index was obtained by calculating the sum of miRNA expression value and Cox coefficients. According to the ranking of prognosis index, users would know the risk group of poor prognosis.

OncoLnc[17]

OncoLnc is an interactive tool for studying survival correlations for lncRNA, miRNA, and mRNA (45, 46). OncoLnc contains patient survival data of 21 cancer types from TCGA mRNAs, miRNAs, and MiTranscriptome data. The users can divide patients into subgroups according to gene expression levels, measure the result between subgroups. OncoLnc allows users to view the results of Kaplan Meier plots of one or multiple types of cancers at one time, provide Cox regression results, and download the full data used in the analysis. It also allows users to explore the survival relevance of inquired genes in 21 types of cancers at one time, this function is helpful to study whether specific genes play important roles in cancer prognosis.

TANRIC[18]

TANRIC is an interactive platform for multiple analysis of lncRNA in cancer (47). It includes the expression profile of lncRNA in more than 6,000 patient samples of 20 cancer types from TCGA and other three independent datasets. TANRIC consists of six modules, users can get the annotation data of lncRNA through module "My lncRNA," and analyze whether lncRNA is related to the survival time of patients (including subtypes prognosis analysis). Users can also use other functions TANRIC to recognize the differential expression of lncRNA in tumor and normal tissue, as well as in tumor subtype or tumor stage, evaluate the differential expression of lncRNA in wild type and gene mutation cancer, evaluate the influence of lncRNA expression on drug sensitivity, and find some signal pathways related to cancer subtype defined by lncRNA.

[17]http://www.oncolnc.org
[18]https://www.tanric.org

Web Servers for Survival Analysis Based on Protein Data

Functional proteomics is a powerful way to understand the pathophysiological mechanism and find the therapeutic target of cancer. In order to find biomarkers for prognosis and targets for treatment improvement, it is necessary to study the correlation between protein and survival. As a part of the Cancer Genome Atlas (TCGA) Project and other works, reverse-phase protein array (RPPA) was used to measure the protein expression in a large number of clinical cancer samples and cell lines (48, 49). This technology provides a necessary condition for the establishment of repeatable prediction model and protein prediction database. Here, we introduce two protein survival analysis databases based on RPPA data (**Table 3**).

TCPAv3.0[19]

TCPAv3.0 is an updated version of TCPA to explore and analyze protein expression based on TCGA RPPA data (50, 51). It integrates protein data and other TCGA data (somatic mutations, SCNAs, DNA methylation, mRNA and miRNA expression, and patient clinical information) and gives comprehensive protein-centric analyses. The users can find protein markers or pathway events that are significantly related to patient survival by using Cox proportional risk model and log rank test. The users can identify which proteins associated with the prognosis of different cancers and subtypes by pan-cancer analysis. The pan-cancer analysis module using multi-omic TCGA data provides researchers a unique way to validate specific protein-driven multi-omic hypotheses in multiple cancer types.

[19]http://tcpaportal.org/

TABLE 4 | Summary of prognosis web servers based on DNA data.

Web server	Datasets	Cancer types	Samples	Data types	Subgroups	Optimal cut-off	Login required
GSCALite	33	33	10,943	Methylation	Yes	No	No
MEXPRESS	33	33	–	Methylation	Yes	No	No
MethSurv	25	25	7,358	Methylation	No	Yes	No
cBioPortal	>100	32	–	Mutation/ CNA	Yes	–	No
GSCALite	33	33	11,124	Mutation	Yes	–	No
CaPSSA	27	26	10,758	Mutation	No	–	No

–, related information is not displayed on the website.

TABLE 5 | Prognostic tools for single type of cancer.

Cancer type	Database	Website	Data type	Reference
Breast cancer	miRpower	http://kmplot.com/mirpower	miRNA	(31)
	BreastMark	http://glados.ucd.ie/BreastMark/index.html	mRNA, miRNA	(60)
	OSbrca	http://bioinfo.henu.edu.cn/BRCA/BRCAList.jsp	mRNA	(19)
Bladder cancer	OSblca	http://bioinfo.henu.edu.cn/BLCA/BLCAList.jsp	mRNA	(17)
Leiomyosarcoma	OSlms	http://bioinfo.henu.edu.cn/LMS/LMSList.jsp	mRNA	(14)
ESCC	OSescc	http://bioinfo.henu.edu.cn/DBList.jsp	mRNA	(15)
KIRC	OSkirc	http://bioinfo.henu.edu.cn/KIRC/KIRCList.jsp	mRNA	(16)
Cervical cancer	OScc	http://bioinfo.henu.edu.cn/CESC/CESCList.jsp	mRNA	(18)
Adrenocortical carcinoma	OSacc	http://bioinfo.henu.edu.cn/ACC/ACCList.jsp	mRNA	(20)
Uveal melanoma	OSuvm	http://bioinfo.henu.edu.cn/UVM/UVMList.jsp	mRNA	(21)
Ovarian cancer	OvMark	http://glados.ucd.ie/OvMark/index.html	mRNA, miRNA	(59)

TRGAted[20]

TRGAted is an intuitive tool for analyzing the correlation between more than 200 proteins and survivals in 31 types of cancers (52). RPPA data (Level 4) contained in TRGAted come from the TCPA Portal. The cancer clinical information provided are comprehensive, including: gender, age, tumor stage, histological type, response to treatment. Users can use Cox proportional hazard model to analyze the prognosis of all proteins in each cancer type, or for a single protein across all cancer types. Comparison with TCPAv3.0, TRGAted provides more survival indicators, and its function of visualizing all proteins in a cancer type can help researchers find survival related proteins in the specific cancer more easily. The users are allowed to download and modify TRGAted for better usability under GPLv3 (GNU General Public License v3.0).

Web Servers for Prognosis Analysis Based on DNA Data

Patients with genetic mutations in tumor cells are more likely to display poor pathological features, resulting in significantly altered overall survival (53). The new generation of sequencing technology has accelerated the study of somatic genetics, identifying patient subgroups with different genomic alteration patterns could facilitate to stratify patients with different clinical outcomes and to propose putative biomarkers. In addition to DNA mutation, DNA methylation is the most studied epigenetic modification which is crucial for facilitating vital biological processes such as embryonic development, genomic imprinting, and X-chromosome inactivation. Aberrant DNA methylation may lead to changes in cellular micro-environment, affect the gene expression pattern, and ultimately result in various pathological conditions including carcinogenesis (54, 55). Several recently developed high-throughput techniques facilitate genome-wide DNA methylation profiling. Some prognostic tools were also developed to facilitate the evaluation of the prognostic properties of CpG methylation data (**Table 4**).

MethSurv[21]

MethSurv is a web tool dedicating for survival analysis based on DNA methylation data including 7,358 samples in 25 different cancer types from TCGA (56). Methsurv provides multiple survival terms analysis, and the home page contains the following modules: single CpG, region based analysis, all cancers, top biomarkers, and gene visualization. Users can retrieve CpG survival analysis results of selected areas of a chromosome, and also search for a gene of interest to explore the survival statistics of all CpGs available. Users can see top biomarkers arranged according to *p*-value of all CpG labeled cancer types in the whole

[20]https://nborcherding.shinyapps.io/TRGAted

[21]https://biit.cs.ut.ee/methsurv/

TABLE 6 | Follow-up information of prognostic web servers.

Web server	OS	DFS	RFS	MFS	PFS	DSS	Others	Total
LOGpc	○	○	○	○	○	○	DFI, PFI, DMFS, DRFS,LMFS, BMFS, EFS	13
GENT2	○	○			○	○		4
PROGgeneV2	○		○	○				3
SurvExpress	○		○	○				3
PRECOG	○					○		2
Oncomine	○							1
PrognoScan	○	○	○		○	○	EFS, DMFS, DRFS	8
KM Plotter	○		○		○	○	DMFS, PPS, FP	7
GSCALite	○							1
UALCAN	○							1
GEPIA	○	○						2
CAS-viewer	○							1
MEXPRESS	○							1
CaPSSA	○	○						2
OncoLnc	○							1
PROGmiRV2	○		○	○				3
SurvMicro	○							1
TANRIC	○							1
TCPAv3.0	○				○			2
TRGAted	○					○	DFI, PFI	4
MethSurv	○							1
cBioPortal	○	○						2

"○", Yes; OS, overall survival; DFS, disease free survival; RFS, relapse free survival; MFS, metastasis free survival; PFS, progression free survival; DSS, disease specific survival; DMFS, distant metastasis free survival; PFI, progression free interval; DFI, disease-free interval; PFI, progression free interval; EFS, event free survival; LMFS, lung metastasis free survival; BMFS, brain metastasis free survival; DRFS, distant relapse free survival; FP, first progression; PPS, post progression survival.

genome. In brief, MethSurv is a valuable platform for preliminary screening of methylation cancer biomarkers.

cBioPortal[22]

cBioPortal provides a visual tool for interactive exploration of multiple cancer genomic datasets (57, 58). It integrates and simplifies the data including somatic mutation, mRNA and microRNA expression, DNA copy-number alterations(CNAs) and methylation, protein, and phosphoprotein RPPA data, so that the users can obtain graphical summaries of large-scale cancer genomic data intuitively. It enables users to inquiry survival analysis based on DNA mutation data and CNA data, the results of OS, and DFS of patients are presented intuitively in the form of Kaplan-Meier plots. Pan-cancer analysis is also allowed.

Prognostic Tools for Single Type of Cancer

Through literature search, 11 prognostic tools for single type of cancer were found (**Table 5**). MiRpower is a part of KMplotter database to analyze the prognostic relevance of miRNAs in breast cancer (31). OSlms, OSescc, OSkirc, OSblca, OScc, OSbrca, OSacc, and OSuvm are bioinformatics tools included in the LOGpc platform for survival analysis of leiomyosarcoma, esophageal squamous cell carcinoma, kidney renal clear cell carcinoma, bladder cancer, cervical cancer, breast cancer, adrenocortical carcinoma, and uveal melanoma (14–21). OvMark and BreastMark are online web servers for prognosis analysis of ovarian cancer and breast cancer, users can detect the prognostic potential of about 17,000 genes and 341 miRNAs in ovarian cancer and breast cancer (59, 60).

[22]http://www.cbioportal.org

DISCUSSION

The development of public databases (such as TCGA and GEO) provides a large number of genomic, epigenomic, transcriptional and proteomic data, and provides the possibility for gene function analysis and biological mechanism discussion (1, 2). The rapid growth of multi-omics data provides more opportunities for the research of cancer molecular mechanism and biological target, but for the researchers without strong computing power and bioinformatics background, they might face many difficulties and challenges in data mining and analysis. Since the EAPC (European Association for Palliative Care) made recommendations for the development of cancer prognostic tools in 2005, a number of prognostic tools have been developed, evolved, and validated (61). In this review, we summarized 22 prognostic bioinformatics tools, which provide survival analysis or with other functions. We analyzed and compared their key information and characteristics, follow-up information for each tool is presented in **Table 6**, strength and limitation are displayed in additional files (**Table S1**). With these tools, researchers can easily explore a large number of

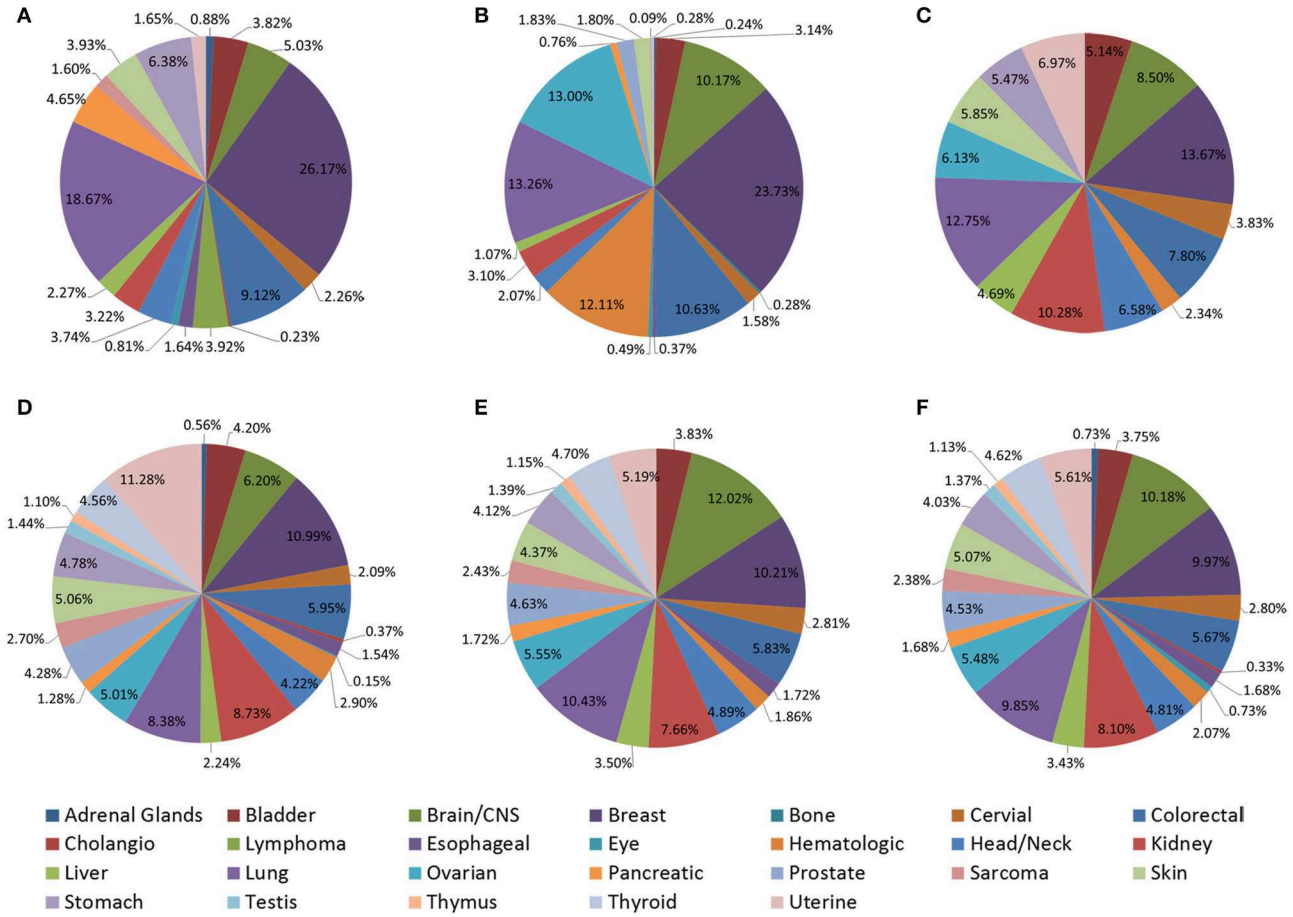


FIGURE 3 | Distribution of cancer types in web servers. **(A)** LOGpc (mRNA level); **(B)** PROGmiRV2 (miRNA level); **(C)** OncoLnc (lncRNA level); **(D)** CaPSSA (mutation level); **(E)** GSCALite (methylation level); **(F)** TCPAv3.0 (protein level).

datasets from complex data platform, find genes, ncRNAs, proteins, gene modifications, or mutations associated with patient survival, ask specific questions and test their hypotheses (48, 62, 63). Comprehensive expression analysis can be carried out by simple clicks, which greatly promotes data mining in research fields, scientific discussions and treatment discovery processes. These tools have the potentials to integrate and personalize the prognostic information for individual patients and provide refined risk estimates for uncertain clinical management scenarios. Meanwhile each database has its own strengths. Some databases focus on survival analysis by collecting datasets of various cancer types, such as LOGpc, PROGgeneV2, KM Plotter, PrognoScan, TRGAted. Some databases provide other functions, UALCAN, and GEPIA have the function of top differential gene display, which provide a way for clinicians and researchers to select possible target genes for diagnosis or treatment, Oncomine, and TCPA provide multidimensional analysis and comparison of datas. GSCALite, TANRIC can be used for drug screening and treatment options by analyzing the correlation between therapeutic targets and lncRNAs. Advances in genome technology and computational biology provide us with an unprecedented opportunity to understand molecular events associated with cancer, and to apply precise cancer treatment. We hope this review will be helpful to clinicians and oncologists who are interested in finding prognostic or predictive features of cancer.

LIMITATION AND PROSPECTIVE

Although these tools provide great convenience for prognostic biomarker development, several key aspects of these prognostic tools remain elusive. Differences in datasets collected and split points may result in significantly different results, so we collected datasets and their source of these web servers (**Figure 3** and **Tables S2–S5**) and found excluding TCGA data, there are significant differences in other data sources. This may be one of the reasons why the analysis results of different tools are not completely consistent. In the future, efforts should be made in data optimization, prognostic tools should be improved to be able to predict multi-gene markers, select optimal cut-off computation, use hierarchical clustering and consider complex multi-omics networks of interactions. In addition more molecular subtypes and clinical information including tumor tissue image and treatment data should be collected and mined to identify

more meaningful prognostic markers through more detailed subtype analysis.

AUTHOR CONTRIBUTIONS

HZ, GZ, LZ, QW, and XG collected data, set up web pages, and drafted the paper. HL, YH, LX, ZY, YL, YA, HD, and WZ contributed to critical revision of the manuscript for intellectual content. All authors edited and approved the final manuscript.

FUNDING

This study was supported by National Natural Science Foundation of China (No. 81602362), Supporting grants of Henan University (No. 2015YBZR048; No. B2015151), Yellow River Scholar Program (No. H2016012), and Program for Innovative Talents of Science and Technology in Henan Province (No. 18HASTIT048), Program for Science and Technology Development in Henan Province (No. 162102310391, No. 172102210187), Program for Scientific and Technological Research of Henan Education Department (No. 14B520022), Program for Young Key Teacher of Henan Province (2016GGJS-214), Kaifeng Science and Technology Major Project (18ZD008), Supporting grant of Bioinformatics Center of Henan University (No. 2018YLJC01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00068/full#supplementary-material

Table S1 | Feature analysis of included prognostic web servers.

Table S2 | Datasets and samples of prognostic web servers based on mRNA.

Table S3 | Datasets and samples of prognostic web servers based on ncRNA.

Table S4 | Datasets and samples of prognostic web servers based on protein.

Table S5 | Datasets and samples of prognostic web servers based on DNA methylation and mutation.

REFERENCES

1. Tomczak K, Czerwinska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol.* (2015) 19:A68–77. doi: 10.5114/wo.2014.47136
2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* (2013) 41:D991–5. doi: 10.1093/nar/gks1193
3. Xu XL, Gong Y, Zhao DP. Elevated PHD2 expression might serve as a valuable biomarker of poor prognosis in lung adenocarcinoma, but no lung squamous cell carcinoma. *Eur Rev Med Pharmacol Sci.* (2018) 22:8731–9. doi: 10.26355/eurrev_201812_16638
4. Sun D, Wang X, Sui G, Chen S, Yu M, Zhang P. Downregulation of miR-374b-5ppromotes chemotherapeutic resistance in pancreatic cancer by upregulating multiple anti-apoptotic proteins. *Int J Oncol.* (2018) 52:1491–503. doi: 10.3892/ijo.2018.4315
5. Yang J, Li A, Li Y, Guo X, Wang M. A novel approach for drug response prediction in cancer cell lines via network representation learning. *Bioinformatics.* (2019) 35:1527–—35. doi: 10.1093/bioinformatics/bty848
6. Hong L, Han Y, Zhang H, Fan D. Prognostic markers in esophageal cancer: from basic research to clinical use. *Expert Rev Gastroenterol Hepatol.* (2015) 9:887–9. doi: 10.1586/17474124.2015.1041507
7. Kang H, Kiess A, Chung CH. Emerging biomarkers in head and neck cancer in the era of genomics. *Nat Rev Clin Oncol.* (2014) 12:11–26. doi: 10.1038/nrclinonc.2014.192
8. Burkhart RA, Ronnekleiv-Kelly SM, Pawlik TM. Personalized therapy in hepatocellular carcinoma: molecular markers of prognosis and therapeutic response. *Surg Oncol.* (2017) 26:138–45. doi: 10.1016/j.suronc.2017.01.009
9. Chou CK, Liu RT, Kang HY. MicroRNA-146b: a novel biomarker and therapeutic target for human papillary thyroid cancer. *Int J Mol Sci.* (2017) 18:636. doi: 10.3390/ijms18030636
10. Gu X, Xue J, Ai L, Sun L, Zhu X, Wang Y, et al. SND1 expression in breast cancer tumors is associated with poor prognosis. *Ann N Y Acad Sci.* (2018) 1433:53–60. doi: 10.1111/nyas.13970
11. Xie L, Dang Y, Guo J, Sun X, Xie T, Zhang L, et al. High KRT8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes.* (2019) 10:E36. doi: 10.3390/genes10010036
12. Szász AM, Lánczky A, Nagy Á, Förster S, Hark K, Green JE, et al. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget.* (2016) 7:49322–33. doi: 10.18632/oncotarget.10337
13. Guerrero-Martínez JA, Reyes JC. High expression of SMARCA4 or SMARCA2 is frequently associated with an opposite prognosis in cancer. *Sci Rep.* (2018) 8:2043. doi: 10.1038/s41598-018-20217-3
14. Wang Q, Xie L, Dang Y, Sun X, Xie T, Guo J, et al. OSlms: a web server to evaluate the prognostic value of genes in leiomyosarcoma. *Front Oncol.* (2019) 9:190. doi: 10.3389/fonc.2019.00190
15. Wang Q, Wang F, Lv J, Xin J, Xie L, Zhu W, et al. Interactive online consensus survival tool for esophageal squamous cell carcinoma prognosis analysis. *Oncol Lett.* (2019) 18:1199–206. doi: 10.3892/ol.2019.10440
16. Xie L, Wang Q, Dang Y, Ge L, Sun X, Li N, et al. OSkirc: a web tool for identifying prognostic biomarkers in kidney renal clear cell carcinoma. *Future Oncol.* (2019) 15:3103–10. doi: 10.2217/fon-2019-0296
17. Zhang G, Wang Q, Yang M, Yuan Q, Dang Y, Sun X, et al. OSblca: a web server for investigating prognostic biomarkers of bladder cancer patients. *Front Oncol.* (2019) 9:466. doi: 10.3389/fonc.2019.00466
18. Wang Q, Zhang L, Yan Z, Xie L, An Y, Li H, et al. OScc: an online survival analysis web server to evaluate the prognostic value of biomarkers in cervical cancer. *Future Oncol.* (2019) 15:3693–9. doi: 10.2217/fon-2019-0412
19. Yan Z, Wang Q, Sun X, Ban B, Lu Z, Dang Y, et al. OSbrca: a web server for breast cancer prognostic biomarker investigation with massive data from tens of cohorts. *Front Oncol.* (2019) 9:1349. doi: 10.3389/fonc.2019.01349
20. Xie L, Wang Q, Nan F, Ge L, Dang Y, Sun X, et al. OSacc: gene expression-based survival analysis web tool for adrenocortical carcinoma. *Cancer Manag Res.* (2019) 11:9145–52. doi: 10.2147/CMAR.S215586
21. Wang F, Wang Q, Li N1, Ge L, Yang M, An Y, et al. OSuvm: an interactive online consensus survival tool for uveal melanoma prognosis analysis. *Mol Carcinog.* (2020) 59:56–61. doi: 10.1002/mc.23128
22. Park SJ, Yoon BH, Kim SK, Kim SY. GENT2: an updated gene expression database for normal and tumor tissues. *BMC Med Genomics.* (2019) 12:101. doi: 10.1186/s12920-019-0514-7
23. Goswami CP, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinform.* (2013) 3:22. doi: 10.1186/2043-9113-3-22
24. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. *BMC Cancer.* (2014) 14:970. doi: 10.1186/1471-2407-14-970
25. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS ONE.* (2013) 8:e74250. doi: 10.1371/journal.pone.0074250
26. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med.* (2015) 21:938–45. doi: 10.1038/nm.3909

27. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia.* (2004) 6:1–6. doi: 10.1016/S1476-5586(04)80047-2
28. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia.* (2007) 9:166–80. doi: 10.1593/neo.07112
29. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics.* (2009) 2:18. doi: 10.1186/1755-8794-2-18
30. Győrffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1809 patients. *Breast Cancer Res Treat.* (2010) 123:725–31. doi: 10.1007/s10549-009-0674-9
31. Nagy Á, Lánczky A, Menyhárt O, Gyorffy B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci Rep.* (2018) 8:9227. doi: 10.1038/s41598-018-27521-y
32. Lánczky A, Nagy Á, Bottai G, Munkácsy G, Szabó A, Santarpia L, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res Treat.* (2016) 160:439–46. doi: 10.1007/s10549-016-4013-7
33. Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. *Bioinformatics.* (2018) 34:3771–2. doi: 10.1093/bioinformatics/bty411
34. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia.* (2017) 19:649–58. doi: 10.1016/j.neo.2017.05.002
35. Tang Z, Li C1, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* (2017) 45:W98–102. doi: 10.1093/nar/gkx247
36. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* (2019) 47:W556–60. doi: 10.1093/nar/gkz430
37. Han S, Kim D, Kim Y, Choi K, Miller JE, Kim D, et al. CAS-viewer: web-based tool for splicing-guided integrative analysis of multi-omics cancer data. *BMC Med Genomics.* (2018) 11:25. doi: 10.1186/s12920-018-0348-8
38. Koch A, Jeschke J, Van Criekinge W, van Engeland M, De Meyer T. MEXPRESS update 2019. *Nucleic Acids Res.* (2019) 47:W561–5. doi: 10.1093/nar/gkz445
39. Jang Y, Seo J, Kim S, Lee S. CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data. *Bioinformatics.* (2019) 35:btz516. doi: 10.1093/bioinformatics/btz516
40. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. *Nat Rev Cancer.* (2018) 18:5–18. doi: 10.1038/nrc.2017.99
41. Zhu Y, Wang J, Wang F, Yan Z, Liu G, Ma Y, et al. Differential microRNA expression profiles as potential biomarkers for pancreatic ductal adenocarcinoma. *Biochemistry.* (2019) 84:575–82. doi: 10.1134/S0006297919050122
42. Li J, Li Z, Zheng W, Li X, Wang Z, Cui Y, et al. LncRNA-ATB: an indispensable cancer-related long noncoding RNA. *Cell Prolif.* (2017) 50:12381. doi: 10.1111/cpr.12381
43. Goswami CP, Nakshatri H. PROGmiR: a tool for identifying prognostic miRNA biomarkers in multiple cancers using publicly available data. *J Clin Bioinform.* (2012) 2:23. doi: 10.1186/2043-9113-2-23
44. Aguirre-Gamboa R, Trevino V. SurvMicro: assessment of miRNA-based prognostic signatures for cancer clinical outcomes by multivariate survival analysis. *Bioinformatics.* (2014) 30:1630–2. doi: 10.1093/bioinformatics/btu087
45. Anaya J. OncoRank: a pan-cancer method of combining survival correlations and its application to mRNAs, miRNAs, and lncRNAs. *Peer J Preprints.* (2016) 4:e2574. doi: 10.7287/peerj.preprints.2574v1
46. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peer J Comput Sci.* (2016) 2:e67. doi: 10.7717/peerj-cs.67
47. Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, et al. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* (2015) 75:3728–37. doi: 10.1158/0008-5472.CAN-15-0273
48. Li J, Akbani R, Zhao W, Lu Y, Weinstein JN, Mills GB, et al. Explore, visualize, and analyze functional cancer proteomic data using the cancer proteome atlas. *Cancer Res.* (2017) 77:e51–4. doi: 10.1158/0008-5472.CAN-17-0369
49. Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin Proteomics.* (2010) 6:129–51. doi: 10.1007/s12014-010-9055-y
50. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. *Nat Methods.* (2013) 10:1046–7. doi: 10.1038/nmeth.2650
51. Chen MM, Li J, Wang Y, Akbani R, Lu Y, Mills GB, et al. TCPA v3.0: an integrative platform to explore the pan-cancer analysis of functional proteomic data. *Mol Cell Proteomics.* (2019) 18:S15–25. doi: 10.1074/mcp.RA118.001260
52. Borcherding N, Bormann NL, Voigt AP, Zhang W. TRGAted: a web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Res.* (2018) 7:1235. doi: 10.12688/f1000research.15789.1
53. Swift SL, Lang SH, White H, Misso K, Kleijnen J, Quek RG. Effect of DNA damage response mutations on prostate cancer prognosis: a systematic review. *Future Oncol.* (2019) 15:3283–303. doi: 10.2217/fon-2019-0298
54. Gyorffy B, Bottai G, Fleischer T, Munkácsy G, Budczies J, Paladini L, et al. Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. *Int J Cancer.* (2016) 138:87–97. doi: 10.1002/ijc.29684
55. Chen X, Zhao C, Zhao Z, Wang H, Fang Z. Specific glioma prognostic subtype distinctions based on DNA methylation patterns. *Front Genet.* (2019) 10:786. doi: 10.3389/fgene.2019.00786
56. Modhukur V, Iljasenko T, Metsalu T, Lokk K, Laisk-Podar T, Vilo J. MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics.* (2018) 10:277–88. doi: 10.2217/epi-2017-0118
57. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* (2013) 6:pl1. doi: 10.1126/scisignal.2004088
58. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095
59. Madden SF, Clarke C, Stordal B, Carey MS, Broaddus R, Gallagher WM, et al. OvMark: a user-friendly system for the identification of prognostic biomarkers in publically available ovarian cancer gene expression datasets. *Mol Cancer.* (2014) 13:241. doi: 10.1186/1476-4598-13-241
60. Madden SF, Clarke C, Gaule P, Aherne ST, O'Donovan N, Clynes M, et al. BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. *Breast Cancer Res.* (2013) 15:R52. doi: 10.1186/bcr3444
61. Simmons CPL, McMillan DC, McWilliams K, Sande TA, Fearon KC, Tuck S, et al. Prognostic tools in patients with advanced cancer: a systematic review. *J Pain Symptom Manage.* (2017) 53:962–70. doi: 10.1016/j.jpainsymman.2016.12.330
62. Deng JL, Xu YH, Wang G. Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Front Genet.* (2019) 10:695. doi: 10.3389/fgene.2019.00695
63. Coebergh van den Braak RRJ, Sieuwerts AM, Kandimalla R, Lalmahomed ZS, Bril SI, van Galen A, et al. High mRNA expression of splice variant SYK short correlates with hepaticdisease progression in chemonaive lymph node negative colon cancer patients. *PLoS ONE.* (2017) 12:e0185607. doi: 10.1371/journal.pone.0185607

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fgene.2020.00045

Identification of Core Gene Expression Signature and Key Pathways in Colorectal Cancer

Xiang Ding, Houyu Duan and Hesheng Luo *

Department of Gastroenterology, Renmin Hospital, Wuhan University, Wuhan, China

Objective: Colorectal cancer (CRC) is considered the most prevalent malignant tumor that contributes to high cancer-related mortality. However, the signaling pathways involved in CRC and CRC-driven genes are largely unknown. We sought to discover a novel biomarker in CRC.

Materials and Methods: All clinical CRC samples (n = 20) were from Renmin Hospital of Wuhan University. We first selected MAD2L1 by integrated bioinformatics analysis of a GSE dataset. Next, the expression of MAD2L1 in tissues and cell lines was verified by quantitative real-time PCR. The effects of MAD2L1 on cell growth, proliferation, the cell cycle, and apoptosis were examined by *in vitro* assays.

Results: We identified 683 shared DEGs (420 upregulated and 263 downregulated), and the top twenty genes (CDK1, CCNA2, TOP2A, PLK1, MAD2L1, AURKA, BUB1B, UBE2C, TPX2, RRM2, KIF11, NCAPG, MELK, NUSAP1, MCM4, RFC4, PTTG1, CHEK1, CEP55, DTL) were selected by integrated analysis. These hub genes were significantly overexpressed in CRC samples and were positively correlated. Our data revealed that the expression of MAD2L1 in CRC tissues is higher than that in normal tissues. MAD2L1 knockdown significantly suppressed CRC cell growth by impairing cell cycle progression and inducing cell apoptosis.

Conclusion: MAD2L1, as a novel oncogenic gene, plays a role in regulating cancer cell growth and apoptosis and could be used as a new biomarker for diagnosis and therapy in CRC.

Keywords: MAD2L1, colorectal cancer, bioinformatics analysis, proliferation, cell cycle, apoptosis

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Xian Shen,
Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, China
Carmelo Laudanna,
Institute for Research in Biomedicine, Spain

***Correspondence:**
Hesheng Luo
xhnk@163.com

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *17 October 2019*
Accepted: *15 January 2020*
Published: *21 February 2020*

Citation:
Ding X, Duan H and Luo H (2020) Identification of Core Gene Expression Signature and Key Pathways in Colorectal Cancer. Front. Genet. 11:45. doi: 10.3389/fgene.2020.00045

INTRODUCTION

Colorectal cancer (CRC) is currently a major public health problem in medicine today. CRC is one of the most frequently occurring malignancies worldwide, with more than 777,000 new cases expected in 2015 and almost 350,000 deaths in developed countries (Ferlay et al., 2015). The risk of developing colorectal cancer depends on different variables that can be classified into lifestyle or behavioral factors and genetically determinant factors. Similar to other cancers, CRC is considered a polyphase disease in which gene distortions, cellular contexts, and environmental influences concur with tumor initiation, progression, and metastasis (Aran et al., 2016). Increasing evidence shows that multiple genes and cellular pathways are involved in the occurrence and development of CRC.

Until now, a lack of knowledge about the exact molecular mechanisms underlying CRC progression has limited the ability to treat advanced disease. On the other hand, so far, the main clinical screening methods for CRC involve endoscopic screening, especially colonoscopy. Colonoscopy has shortcomings such as poor patient compliance, the influence of family history, inconvenience, and high cost and risk. Therefore, it is of great significance to understand the molecular mechanisms of CRC proliferation, apoptosis and invasion in order to develop more effective diagnostic and therapeutic strategies.

The recently adopted high-throughput gene microarray analysis of tumors and samples from patients and healthy people allows us to share and explore global molecular tumors at different levels of the landscape from somatic mutations and copy number changes to genome-level gene expression at the transcriptome level, as well as epigenetic changes (Liu et al., 2017; Sun et al., 2017; Chen et al., 2017). In this study, we downloaded the GSE117606 dataset from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) database using R software for the comprehensive identification of differentially expressed genes (DEGs). Then, we established a protein–protein interaction (PPI) network of DEGs to screen out the first 20 hub genes with a high degree of connectivity. In addition, we also analyzed Gene Ontology involving the biological processes (BPs), molecular functions (MFs), and cellular components (CCs) of the DEGs as well as their KEGG pathways. The potential correlation and expression levels were analyzed *via* Gene Expression Profiling Interactive Analysis (GEPIA) (http://gepia.cancer-pku.cn/index.html).

Our data showed that the expression of MAD2L1 is significantly higher in CRC tissues than in normal tissues. The cell cycle progression could be slowed, and apoptosis could be induced by knocking down MAD2L1, which directly leads to the inhibition of the growth of CRC cells. In conclusion, MAD2L1 can be used as a new diagnostic indicator and guide the combined treatment of CRC.

MATERIALS AND METHODS

Microarray Data

We downloaded the gene expression profile of GSE117606 from the GEO database, a free public database. The GSE117606 dataset has a total of 208 samples, containing 74 CRC samples and 65 normal colon tissues and was based on the Agilent GPL25373 platform (HT_HG-U133_Plus_PM) Affymetrix HT HG-U133+ PM Array Plate (CDF: HTHGU133Plus PM_Hs_ENTREZG_20) by Joke Reumers et al. We also downloaded the Series Matrix File of GSE117606 from the GEO database.

Data Preprocessing

The expression values of all probes in each sample were reduced to a single value by determining the mean expression value *via* the aggregate function method (Li, 1991). Missing data were assigned using the k-nearest neighbor method (Altman, 1992). Quantile normalization for complete data was performed using the preprocessCore package in Bioconductor (Bolstad et al., 2003). When many probes were mapped to a gene, the median of the data was defined as the level of expression of that gene. However, when many genes were located by a probe, the probe was considered to lack specificity and was removed from the analysis.

Identification of DEGs

We utilized the "limma" R package (Ritchie et al., 2015) to identify the DEGs between CRC samples and normal ovarian samples. Adjusted $P < 0.05$ and $|\log \text{fold change (FC)}| > 1$ were chosen as the cutoff criteria. The adjusted *P*-value (adj. *P*) was applied to help correct false positives. The heat map and volcano plot were drawn with the "gplots" package in R 3.5.3 (Galili et al., 2018).

A total of 683 DEGs were found, including 420 upregulated genes and 263 downregulated genes, and we selected the top 20 genes with a high degree of connectivity as hub genes.

Gene Ontology and KEGG Pathway Analysis of DEGs

Gene Ontology (GO) analysis can be used to annotate genes and their products with cellular components (CCs), molecular functions (MFs), biological pathways (BPs), and other functions (Gaudet et al., 2017). The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases that address genomic and biological pathways related to diseases and drugs. KEGG is essentially a resource for the comprehensive understanding of biological systems and some high-level genomic functional information (Kanehisa, 2002). Database for Annotation, Visualization, and Integrated Discovery (DAVID, http://david.ncifcrf.gov) (version 6.8) is an online biological information database that integrates a large amount of biological data and related analysis tools, providing systematic and comprehensive biological function annotation information for high-throughput gene expression (Huang et al., 2007). $P < 0.05$ was used as the cut-off criterion for statistically significant differences. To visualize the key molecular functions, biological processes, cellular components, and KEGG pathways of the DEGs, the DAVID online database was used to perform biological analysis.

PPI Network and Module Analysis

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is an online tool that was designed to evaluate and integrate protein–protein interaction (PPI) information, such as physical and functional associations. To date, a total of 9,643,763 proteins from 2,031 organisms have been covered in STRING version 11.0 (Szklarczyk et al., 2015). To evaluate the interrelationships among these DEGs, we first drew the network of DEGs in STRING and then visualized the PPI network by using Cytoscape software. Moreover, we set the maximum number of interacting bodies to 0 and used a confidence score of 0.7 as the

cut-off criterion. Additionally, the Molecular Complex Detection (MCODE) app was also employed to select modules of the PPI network in Cytoscape according to node score cut-off = 0–2, degree cut-off = 2, max.depth = 100, and k–core = 2. With DAVID, the gene pathways of the three modules were analyzed. Additionally, 20 hub genes were mapped into STRING according to a confidence score ≥0.4 and a maximum number of interactors ≤5. We also used GO and KEGG pathway analysis to investigate their underlying information.

Comparison of the Hub Genes' Expression Levels

GEPIA (http://gepia.cancer-pku.cn/index.html) is a newly developed interactive web server designed by Zefang Tang, Chenwei Li, and Boxi Kang of the Zhang Lab, Peking University, designed to analyze the RNA sequence expression data of 9,736 tumors and 8,587 normal samples from the TCGA and GTEx projects using a standard processing pipeline. GEPIA provides customizable capabilities, such as tumor/normal differential expression analysis, profiling by cancer type or pathological stages, patient survival analysis, similar gene testing, correlation analysis, and dimensional reduction analysis (Tang et al., 2017). In our study, we mainly used boxplots to visualize hub gene expression in CRC and normal colon tissues. Then, we analyzed the top 20 hub genes' correlation with a scatter plot. The Human Protein Atlas (HPA, https://www.proteinatlas.org/) is a Swedish-based program initiated in 2003 with the aim of mapping all human proteins in cells, tissues, and organs using the integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology (Uhlen et al., 2017). We further verified the expression of MAD2L1 by obtaining immunohistochemical data based on the HPA in patients with or without CRC.

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method for exploring whether a given gene set is significantly enriched in a group of gene markers ranked by their relevance with a phenotype of interest. The curated KEGG pathway V5.2 data set was used to compare the impaired pathways in normal and colon cancer samples. In addition, the gene sets with fewer than 15 genes or more than 500 genes were excluded. The phenotype label was set as colon cancer *versus* control. The t-statistic mean of the genes was computed in each KEGG pathway using a permutation test with 1,000 replications. The upregulated pathways were defined by a normalized enrichment score (NES) > 0, and the downregulated pathways were defined by an NES <0. Pathways with an FDR P value ≤1 were considered significantly enriched.

Validation Based on CRC Clinical Samples

To further verify the data from GEO, we conducted quantitative real-time PCR (qRT-PCR) to quantify the expression level of MAD2L1 in clinical CRC patient samples (n = 20) from Renmin Hospital of Wuhan University (Wuhan, China). Written informed consent was obtained from all patients. This study was approved by the Institute Research Ethics Committee of Renmin Hospital of Wuhan University.

Cell Lines and Cell Transfection

All cell lines, including the normal cell line NCM460 and the CRC cell lines HT-29, HCT116, SW620, and SW480, were purchased from Bioyear Biotechnology. The cells were cultured in RPMI-1640 medium supplemented with 10% FBS (Thermo Fisher Scientific). All cells were maintained in a humidified incubator with 5% CO_2 at 37°C. A total of 1×10^4 cells/ml were plated approximately 24 h before transfection. Once the cells reached 40%–60% confluence in each well of a 96-well plate, the cells were transfected with 2.5 nM siRNA/NC (RiboBio, Guangzhou, China) using Lipofectamine 2000 (Thermo Fisher Scientific) at the indicated concentrations according to the manufacturer's instructions. Six hours later, the culture medium was replaced with fresh medium containing 10% FBS. The cells were harvested after 24 h of transfection for the following assays.

The siRNA sequences were as follows:

Si-h-MAD2L1: forward, 5'-GGGUCCAAAGUUGAGUGAGUCUUGAdTdT-3'; reverse, 5'-CGGACUCACCUUGCUUGUAACUACUdTdT-3'.

RNA Extraction, Reverse Transcription (RT)-PCR, and qRT-PCR

Total RNA was extracted from cells using TRIzol reagent (Invitrogen™). Reverse-transcribed complementary DNA was synthesized using the PrimeScript™RT Reagent Kit (Takara). The RT-PCR conditions were 37°C for 15 min, 85°C for 5 s, and held at 4°C. After the dilution (1:4) of cDNA with nuclease-free water, qRT-PCR was performed by a StepOne™ Real-Time PCR system and SYBR® Premix Ex Taq™. The mixes were predenatured at 95°C for 1 min, followed by 40 cycles of denaturation at 95°C for 15 s and 72°C for 45 s. The results were normalized to GAPDH expression. The relative expression level of MAD2L1 was calculated by the $2^{-\Delta\Delta Ct}$ method.

The primers used for qRT-PCR were as follows: GAPDH forward, 5'-CATCATCCCTGCCTCTACTGG-3'; and reverse, 5'-GTGGGTGTCGCTGTTGAAGTC-3'; MAD2L1 forward, 5'-GCAAAAGATGACAGTGCACCC-3'; and reverse, 5'-GTGGTCCCGACTCTTCCCAT -3'.

Colony Formation Assay

Twenty-four hours after SW620 cells were infected with siRNA, approximately 300 cells were seeded on each well of a six-well plate. The cells were allowed to incubate at 37°C for 14 days. Then, the cells were fixed, stained with crystal violet, and photographed. ImageJ software (1.48 u; National Institutes of Health) was used to count the number of clones per well.

Cell Cycle Analysis

Twenty-four hours after siRNA interference, SW620 cells were harvested, centrifuged, and resuspended in 1× PBS. The cells

were fixed in 70% ethanol overnight. On the second day, after being washed with 1× PBS solution and centrifuged, the cells were resuspended in 1× PBS solution and incubated with RNaseA at 37°C for 30 min. Finally, the cells were stained with propidium iodide and analyzed by a FACSCalibur system (BD Biosciences).

Apoptosis Analysis

SW620 cells were transfected with siRNA for 24 h, harvested, and centrifuged. Then, the supernatant was removed and resuspended in 1× PBS solution. This procedure was repeated three times with 1×10^6 cells per well, and then the cells were stained with an Annexin V/FITC and PI kit. After staining, the cells were analyzed with a FACSCalibur system (BD Biosciences).

Statistical Analysis

All experiments were performed at least three times, and each independent test was carried out in triplicate for each condition under the protocol and according to the manufacturer's instructions. All statistical analyses were performed using PASW Statistics 19.0 (IBM) or GraphPad Prism 6 software (GraphPad Software, Inc.).

RESULTS

Identification of DEGs and Hub Genes

A total of 74 CRC samples and 65 normal samples were analyzed. The series from each chip was analyzed separately using R software, and finally, the DEGs, using adjusted P value < 0.05 and logFC ≥ 1 or logFC ≤ −1 as the cut-off criteria, were identified. A total of 683 DEGs were identified after analyzing GSE117606, 420 of which were upregulated genes, and 263 were downregulated (**Figure 1B**). **Figure 1A** shows the performance level of the DEGs with a fold change of 1. In addition, 20 hub genes were identified from high to low according to their degree of connectivity (**Table 1**).

GO Function and KEGG Pathway Enrichment Analysis

To obtain a more comprehensive and in-depth understanding of the selected DEGs, we analyzed the GO function and KEGG pathway enrichment by DAVID. After importing all DEGs into DAVID, we discovered the functions of the upregulated DEGs and downregulated DEGs by GO analysis. More specifically, these DEGs were mainly enriched in biological processes (BPs)

TABLE 1 | Top 20 hub genes with higher degree of connectivity.

Gene	Degree of connectivity	Adjusted p value
CDK1	55	4.64E-50
CCNA2	46	8.09E-28
TOP2A	41	3.81E-25
PLK1	40	3.35E-22
MAD2L1	39	2.61E-21
AURKA	38	1.39E-30
BUB1B	37	3.23E-38
UBE2C	37	2.11E-28
TPX2	36	1.57E-33
RRM2	36	2.23E-22
KIF11	35	4.54E-31
NCAPG	34	2.35E-27
MELK	34	1.01E-25
NUSAP1	33	3.24E-28
MCM4	29	2.76E-26
RFC4	29	3.04E-22
PTTG1	29	1.78E-24
CHEK1	29	2.05E-37
CEP55	29	1.66E-24
DTL	28	5.48E-25

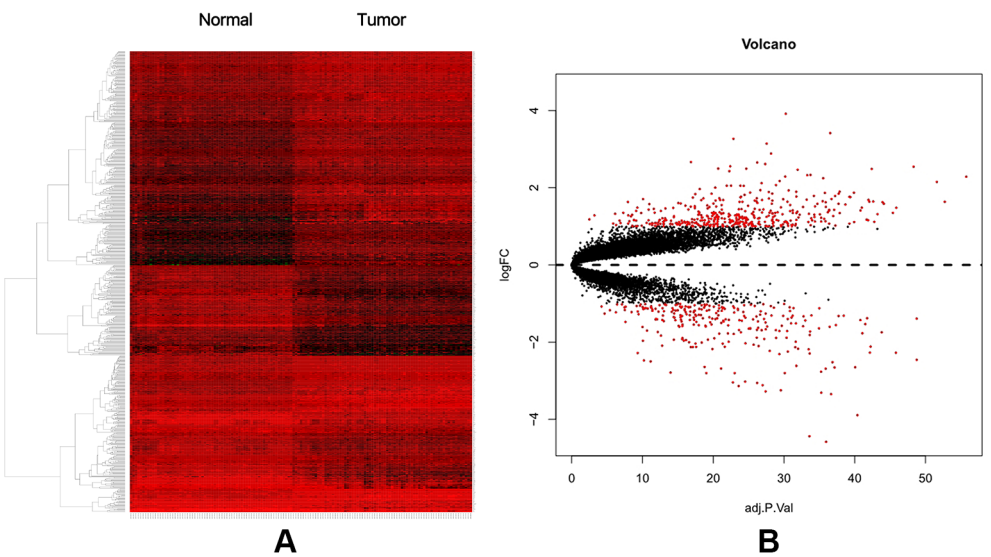


FIGURE 1 | **(A)** Heat map of DEGs. **(B)** Volcano plot of genes detected in CRC. Red means upregulated and downregulated DEGs; black means no difference.

involving collagen catabolic process, extracellular matrix organization, collagen fibril organization, cell division, and G1/S transition of the mitotic cell cycle for the upregulated genes; and bicarbonate transport, muscle contraction, regulation of intracellular pH, chloride transmembrane transport, and one-carbon metabolic process for the downregulated genes. Regarding function (MF), the DEGs were involved in extracellular matrix structural constituent, extracellular matrix binding, platelet-derived growth factor binding, chemokine activity, and calcium ion binding for the upregulated genes; and chloride channel activity, carbonate dehydratase activity, NAD binding, hormone activity, and intracellular calcium activated chloride channel activity for the downregulated genes. In addition, GO cell component (CC) analysis revealed that the upregulated DEGs were principally enriched in the proteinaceous extracellular matrix, extracellular region, extracellular space, collagen trimer, and extracellular matrix, while the downregulated DEGs were mainly enriched in extracellular exosomes, extracellular space, integral components of the plasma membrane, brush border membrane, and apical plasma membrane (**Table 2**).

Table 3 shows the most significantly enriched KEGG pathways of the upregulated and downregulated DEGs. The upregulated DEGs were enriched in the cell cycle, ECM-receptor interaction, focal adhesion, protein digestion and absorption, and the PI3K-Akt signaling pathway, while the downregulated DEGs were enriched in mineral absorption, proximal tubule bicarbonate reclamation, retinol metabolism, pentose and glucuronate interconversions, and steroid hormone biosynthesis. **Figures 2A–C** present a GO and KEGG pathway enrichment plot of CRC.

Hub Genes and Module Screening of the PPI Network

Based on querying STRING protein information from the public database, we constructed a PPI network of the top 20 hub genes according to the degree of connectivity (**Figure 2D**). The top 20 hub genes with a high degree of connectivity were as follows: CDK1, CCNA2, TOP2A, PLK1, MAD2L1, AURKA, BUB1B, UBE2C, TPX2, RRM2, KIF11, NCAPG, MELK, NUSAP1, MCM4, RFC4, PTTG1, CHEK1, CEP55, and DTL. Based on the GO function and KEGG pathway analysis, we found that CDK1, MAD2L1, PLK1, BUB1B, CHEK1, PTTG1, CCNA2, and MCM4 were enriched in the cell cycle. To detect the most important module in this PPI network, we used the MCODE plug-in. The top 3 modules were selected (**Figure 3**). KEGG pathway analysis revealed that the top 3 modules were mainly associated with the cell cycle, ribosome biogenesis in eukaryotes, and the chemokine signaling pathway (**Table 4**).

TABLE 2 | Gene Ontology analysis of differentially expressed genes associated with colorectal cancer.

Expression	Category	Term	Count	%	*P* value	FRD
Upregulated	GOTERM_BP_DIRECT	GO:0030574~collagen catabolic process	19	3.38	6.88E-16	1.14E-12
	GOTERM_BP_DIRECT	GO:0030198~extracellular matrix organization	25	4.45	4.39E-12	7.53E-09
	GOTERM_BP_DIRECT	GO:0030199~collagen fibril organization	13	2.31	1.60E-11	2.75E-08
	GOTERM_BP_DIRECT	GO:0051301~cell division	26	4.63	1.28E-07	2.19E-04
	GOTERM_BP_DIRECT	GO:0000082~G1/S transition of mitotic cell cycle	14	2.49	2.40E-07	4.11E-04
	GOTERM_CC_DIRECT	GO:0005201~extracellular matrix structural constituent	12	2.14	1.18E-07	1.73E-04
	GOTERM_CC_DIRECT	GO:0050840~extracellular matrix binding	8	1.42	7.27E-07	1.06E-03
	GOTERM_CC_DIRECT	GO:0048407~platelet-derived growth factor binding	6	1.07	1.55E-06	2.26 E-03
	GOTERM_CC_DIRECT	GO:0008009~chemokine activity	9	1.60	6.77E-06	9.87 E-03
	GOTERM_CC_DIRECT	GO:0005509~calcium ion binding	31	5.52	2.18E-04	0.32
	GOTERM_MF_DIRECT	GO:0005578~proteinaceous extracellular matrix	28	4.98	4.16E-12	5.63E-09
	GOTERM_MF_DIRECT	GO:0005576~extracellular region	71	12.63	2.09E-10	2.83E-07
	GOTERM_MF_DIRECT	GO:0005615~extracellular space	61	10.85	2.12E-09	2.87E-06
	GOTERM_MF_DIRECT	GO:0005581~collagen trimer	14	2.49	3.07E-08	4.16E-05
	GOTERM_MF_DIRECT	GO:0031012~extracellular matrix	22	3.91	4.80E-07	6.49E-04
Downregulated	GOTERM_BP_DIRECT	GO:0015701~bicarbonate transport	8	2.60	1.19E-06	1.91 E-03
	GOTERM_BP_DIRECT	GO:0006936~muscle contraction	10	3.25	8.46E-06	0.01
	GOTERM_BP_DIRECT	GO:0051453~regulation of intracellular pH	6	1.95	8.39E-05	0.13
	GOTERM_BP_DIRECT	GO:1902476~chloride transmembrane transport	8	2.60	1.75E-04	0.28
	GOTERM_BP_DIRECT	GO:0006730~one-carbon metabolic process	5	1.63	5.23E-04	0.83
	GOTERM_CC_DIRECT	GO:0005254~chloride channel activity	6	1.95	4.91E-04	6.73
	GOTERM_CC_DIRECT	GO:0004089~carbonate dehydratase activity	4	1.30	5.82E-04	7.98
	GOTERM_CC_DIRECT	GO:0051287~NAD binding	5	1.63	1.13 E-03	1.54
	GOTERM_CC_DIRECT	GO:0005179~hormone activity	6	1.95	5.77 E-03	7.65
	GOTERM_CC_DIRECT	GO:0005229~intracellular calcium activated chloride channel activity	3	9.75	0.02	19.61
	GOTERM_MF_DIRECT	GO:0070062~extracellular exosome	69	2.24	1.68E-08	2.13E-05
	GOTERM_MF_DIRECT	GO:0005615~extracellular space	38	12.35	3.90E-06	4.95 E-03
	GOTERM_MF_DIRECT	GO:0005887~integral component of plasma membrane	39	12.68	4.80E-06	6.09 E-03
	GOTERM_MF_DIRECT	GO:0031526~brush border membrane	7	2.28	3.85E-05	0.05
	GOTERM_MF_DIRECT	GO:0016324~apical plasma membrane	13	4.23	2.91E-04	0.37

TABLE 3 | KEGG pathway analysis of differentially expressed genes associated with colorectal cancer.

Category	Term	Count	%	P value	Genes	FRD
Upregulated	hsa04110: Cell cycle	15	0.03	1.06E-06	CDK1, DBF4, SKP2, CHEK1, PTTG1, MCM4, WEE1, YWHAG, CCND1, MAD2L1, MCM7, PLK1, PCNA, BUB1B, CCNA2	0.00133
	hsa04512: ECM-receptor interaction	12	0.02	5.25E-06	COL4A1, ITGAV, COMP, COL3A1, COL1A2, ITGA2, COL1A1, COL11A1, THBS2, COL5A2, COL5A1, SPP1	0.00658
	hsa04510: Focal adhesion	16	0.03	9.37E-05	COL4A1, COL3A1, MET, ITGA2, COL5A2, COL5A1, CCND1, ITGAV, COMP, VEGFA, COL1A2, PDGFRB, COL1A1, THBS2, COL11A1, SPP1	0.11741
	hsa04974: Protein digestion and absorption	10	0.02	2.07E-04	COL4A1, COL7A1, COL3A1, COL1A2, COL12A1, COL1A1, COL11A1, COL5A2, COL5A1, COL10A1	0.25962
	hsa04151: PI3K-Akt signaling pathway	19	0.03	1.19 E-03	COL4A1, COL3A1, MET, ITGA2, COL5A2, COL5A1, DDIT4, YWHAG, EIF4EBP1, CCND1, ITGAV, COMP, VEGFA, COL1A2, PDGFRB, COL1A1, THBS2, COL11A1, SPP1	1.48049
Downregulated	hsa04978:Mineral absorption	8	0.03	4.05E-06	SLC26A3, TRPM6, CLCN2, MT1M, SLC9A3, MT1E, ATP1A2, MT1F	0.00484
	hsa04964:Proximal tubule bicarbonate reclamation	6	0.02	2.13E-05	SLC9A3, CA4, ATP1A2, CA2, SLC4A4, PCK1	0.02546
	hsa00830:Retinol metabolism	7	0.02	4.22E-04	ALDH1A1, UGT1A6, UGT2B17, ADH1C, DHRS9, ADH1B, UGT2B28	0.50270
	hsa00040:Pentose and glucuronate interconversions	5	0.02	1.55 E-03	UGT1A6, UGT2B17, AKR1B10, UGDH, UGT2B28	1.82991
	hsa00140:Steroid hormone biosynthesis	6	0.02	1.89 E-03	HSD3B2, UGT1A6, UGT2B17, HSD17B2, HSD11B2, UGT2B28	2.23810

KEGG, Kyoto Encyclopedia of Genes and Genomes; FDR, false discovery rate.

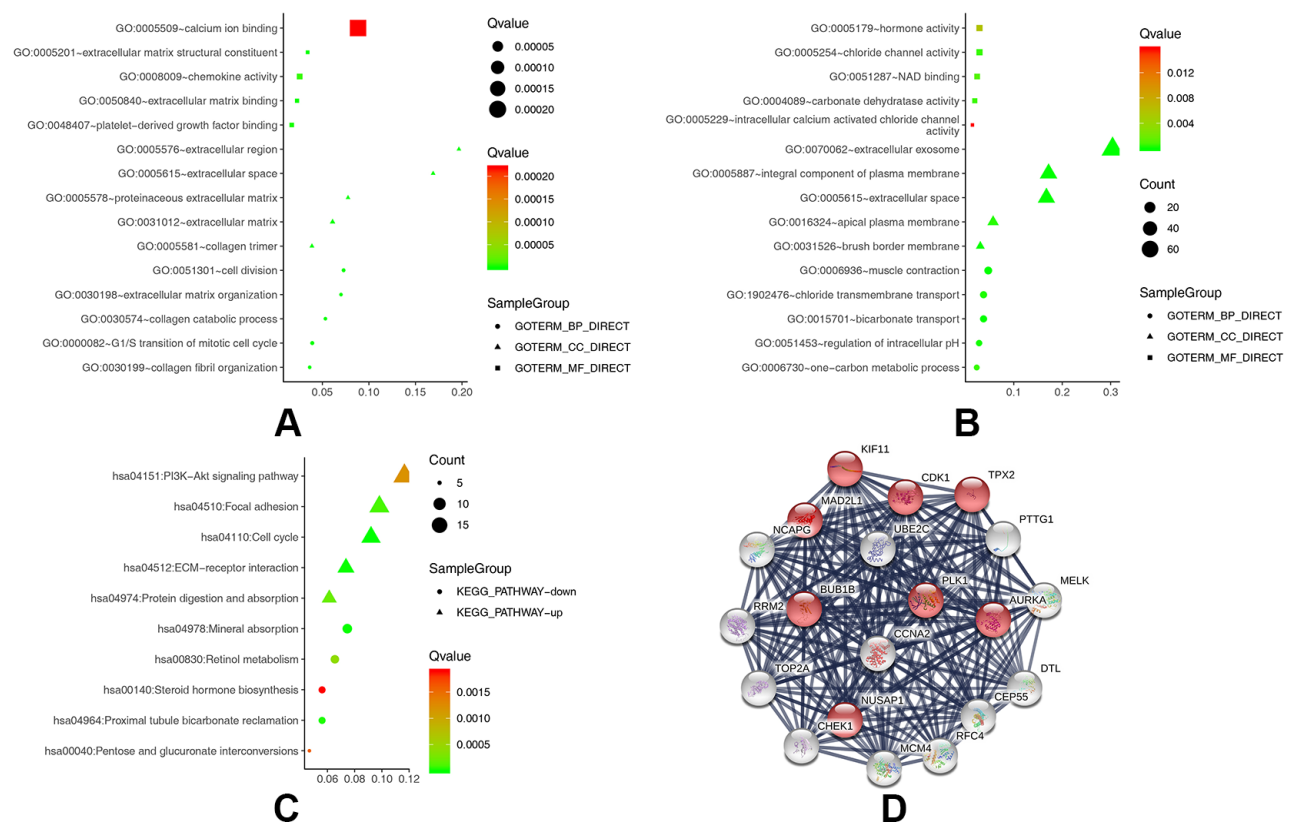


FIGURE 2 | (A) GO analysis of upregulated DEGs. **(B)** GO analysis of downregulated DEGs. **(C)** KEGG pathway of DEGs. **(D)** The protein–protein interaction (PPI) network of the top 20 hub genes.

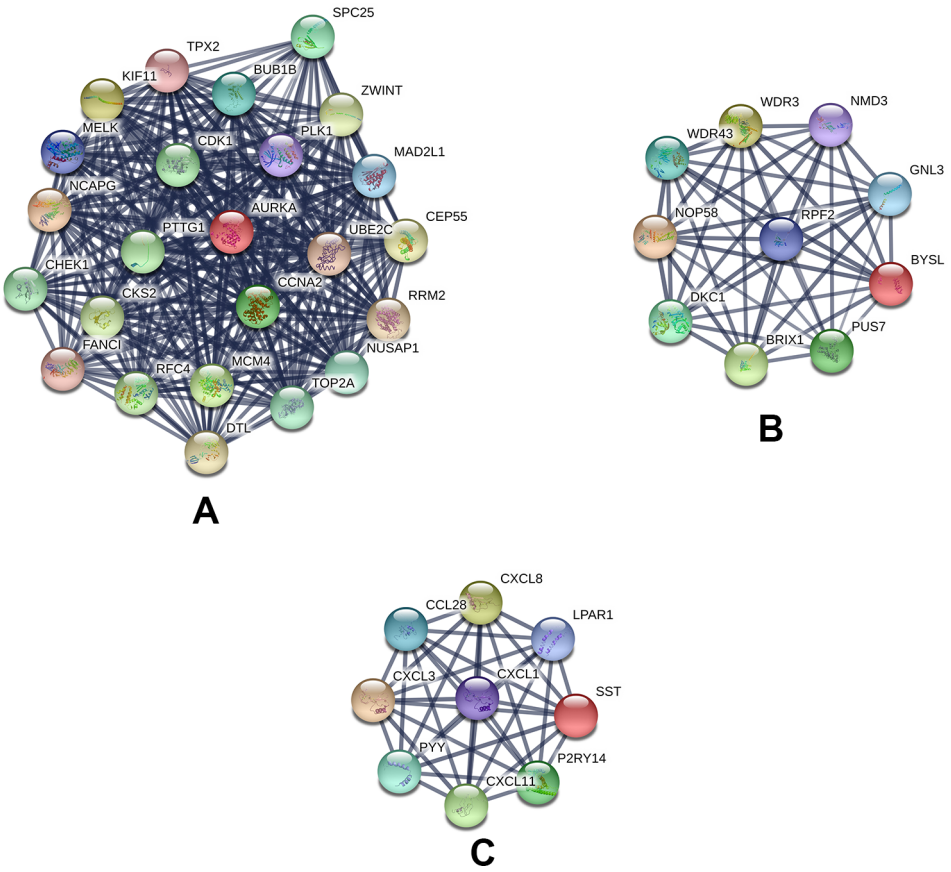


FIGURE 3 | Top 3 modules from the protein–protein interaction network: **(A)** module 1, **(B)** module 2, **(C)** module 3.

TABLE 4 | The enriched pathways of top 3 modules.

Category	Term	Count	%	P value	Genes	FRD
Module 1	hsa04110:Cell cycle	8	33.33	8.19E-10	CDK1, MAD2L1, PLK1, BUB1B, CHEK1, PTTG1, CCNA2, MCM4	6.49E-07
	hsa04114:Oocyte meiosis	5	20.83	4.10E-05	CDK1, MAD2L1, PLK1, AURKA, PTTG1	0.03253
	hsa04914:Progesterone-mediated oocyte maturation	4	16.67	5.10E-04	CDK1, MAD2L1, PLK1, CCNA2	0.40350
	hsa04115:p53 signaling pathway	3	12.5	6.80 E-03	CDK1, RRM2, CHEK1	5.26615
	hsa05166:HTLV-I infection	4	16.67	0.01	MAD2L1, BUB1B, CHEK1, PTTG1	8.25324
Module 2	hsa03008:Ribosome biogenesis in eukaryotes	6	60	2.88E-10	DKC1, WDR3, NOP58, WDR43, NMD3, GNL3	6.51E-08
Module 3	hsa04062:Chemokine signaling pathway	5	0.53	1.70E-05	CXCL1, CXCL3, CXCL8, CXCL11, CCL28	0.01435
	hsa04060:Cytokine-cytokine receptor interaction	5	0.53	4.89E-05	CXCL1, CXCL3, CXCL8, CXCL11, CCL28	0.04128
	hsa05134:Legionellosis	3	0.32	1.24 E-03	CXCL1, CXCL3, CXCL8	1.04102
	hsa05132:Salmonella infection	3	0.32	2.90 E-03	CXCL1, CXCL3, CXCL8	2.42586
	hsa04621:NOD-like receptor signaling pathway	2	0.21	5.56 E-02	CXCL1, CXCL8	38.32740

The Expression Level and Correlation Analyses of the Twenty Hub Genes in GEPIA

GEPIA is an interactive online server for exploring large data sets from the TCGA and GTEx projects. To confirm the reliability of the twenty identified hub genes from the data sets, we used GEPIA to verify the correlation between them, and they were obviously positively correlated with each other in CRC (**Figure 4A**). GEPIA was also used to determine the expression levels of the top ten genes in CRC. **Figure 4B** shows that these genes were all significantly overexpressed in the colon cancer (COAD) samples compared to the normal samples.

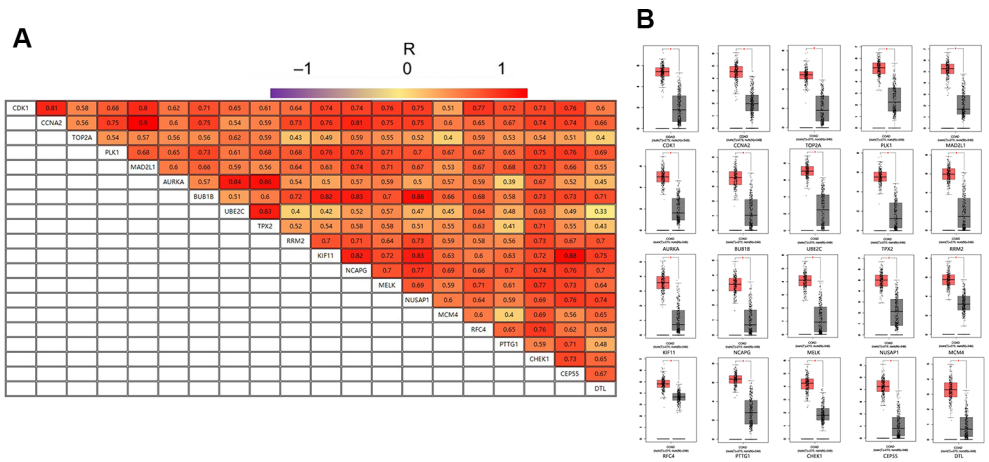


FIGURE 4 | (A) The correlation analysis of the 20 hub genes. **(B)** Expression levels of the 20 hub genes in CRC compared to the normal samples. Notes: R is the Pearson correlation coefficient. Abbreviations: CRC, colorectal cancer.

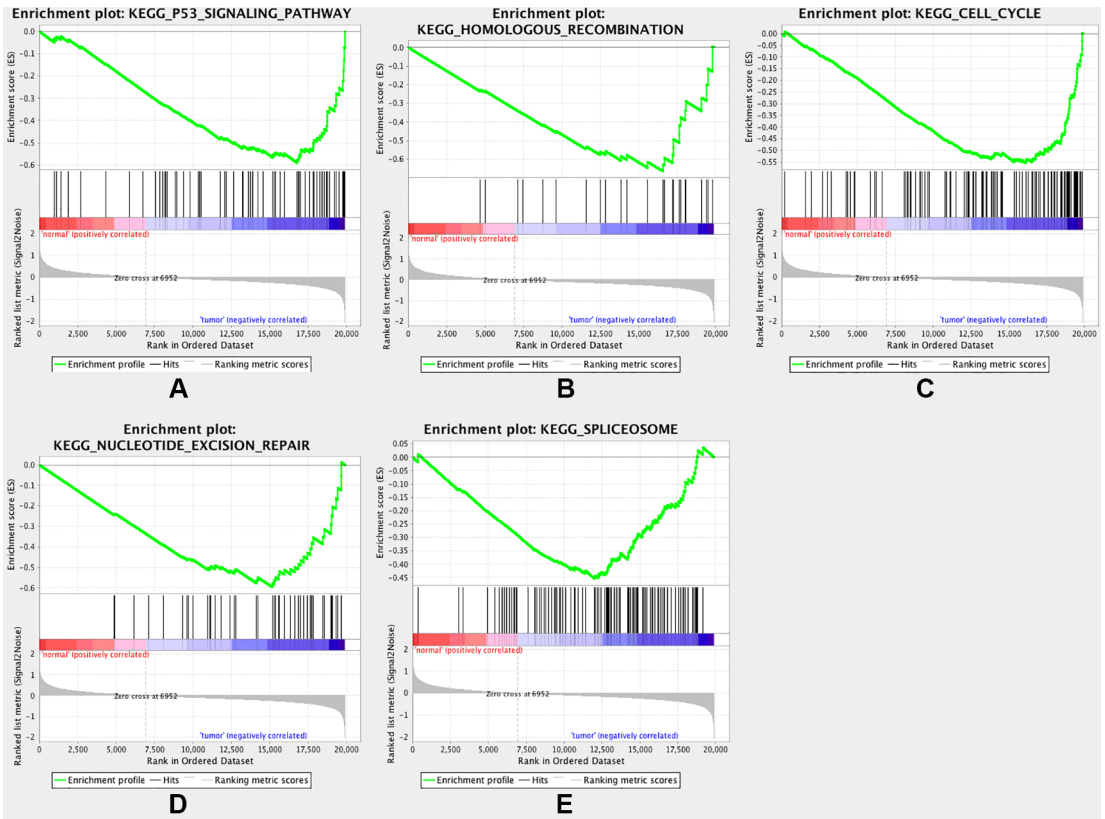


FIGURE 5 | Gene set enrichment analysis (GSEA). Listed pictures are five representative functional gene sets enriched in CRC with MAD2L1 highly expressed.

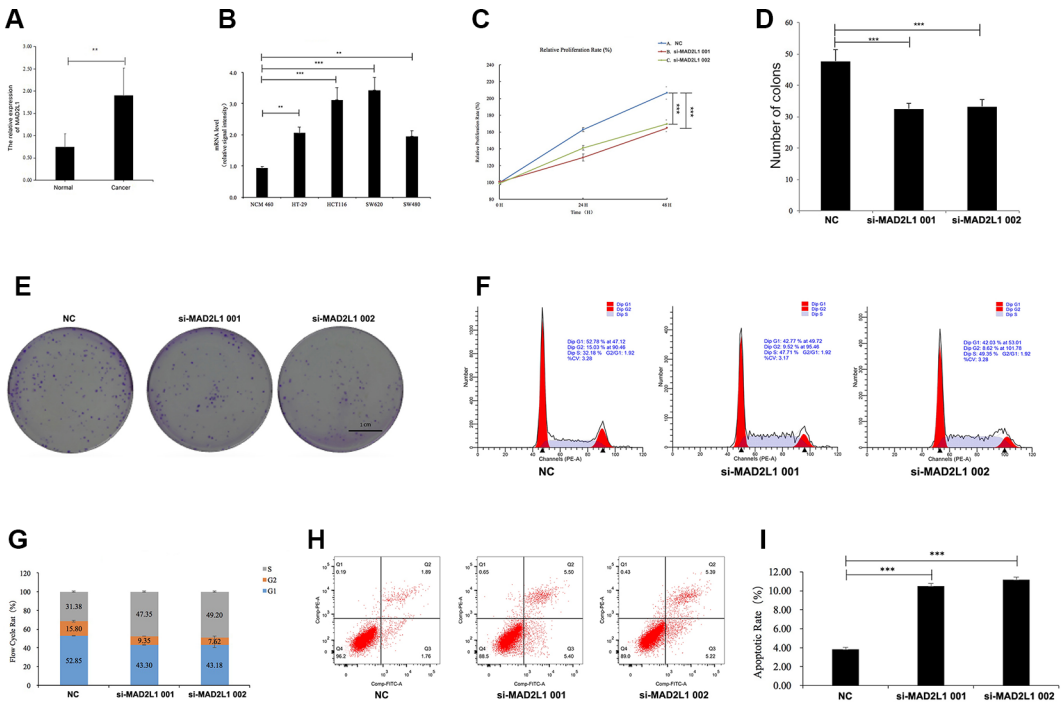


FIGURE 6 | MAD2L1 knockdown suppressed colon cancer cell proliferation by impairing cell cycle progression and inducing apoptosis. Notes: **(A)** Expression level of MAD2L1 gene in 20 paired CRC tissues (n = 3; **P < 0.01; two-tailed t-test). **(B)** Expression level of MAD2L1 gene in colon normal cell line NCM460 and CRC cell line HT-29, HCT116, SW620 and SW480 (n = 3; **P < 0.01, ***P < 0.001; two-tailed t-test). **(C)** The cell proliferation rate was analyzed by CCK-8 assay. All value were mean ± SD (n = 3; ***P < 0.001; two-tailed t-test). **(D)**, **(E)** Colony formation assays were performed (n = 3; ***P < 0.001; two-tailed t-test). **(F, G)** Distribution of cells in three cell cycle phases was examined by flow cytometry assay, and the graph shows quantification for each phase. **(H)** For measurement of apoptotic cells, cells were stained with both AV and PI and analyzed by an image flow assay. **(I)** Graph illustrating the quantification of apoptotic cells (n = 3; ***P < 0.001; two-tailed t-test). Abbreviations: AV, Annexin V FITC; CCK-8, cell counting kit-8, PI, propidium iodide; NC, negative control.

Gene Set Enrichment Analysis

To gain further insight into the functions of the DEGs, GSEA was conducted to map the DEGs into the KEGG pathway database. Under the cut-off criteria of FDR <0.05, |enrichment score (ES)| > 0.6, and gene size ≥100, the top five pathways were "p53 signaling pathway," "homologous recombination," "cell cycle," "nucleotide excision repair", and "spliceosome" (**Figure 5**).

Expression Patterns of MAD2L1 in CRC.

To identify the expression level of MAD2L1 in CRC, we performed qRT-PCR to confirm the expression of MAD2L1 in 20 paired clinical samples, in which the mean expression level of MAD2L1 was notably higher in CRC tissues than in normal tissues (**Figure 6A**). Next, we measured the expression of MAD2L1 in various cell lines, including the normal cell line NCM460 and the CRC cell lines HT-29, HCT116, SW620, and SW480. The expression of MAD2L1 was higher in tumor cells than in normal cells (**Figure 6B**), which is similar to the results from the four datasets in GEO and the GEPIA results, suggesting that our results for these genes are reliable.

Knockdown of MAD2L1 Suppressed Cell Growth by Impairing Cell Cycle Progression and Inducing Cell Apoptosis

To determine whether MAD2L1 could be a therapeutic target in CRC, we inactivated MAD2L1 by using siRNAs in SW620 cells. We found that the MAD2L1 knockdown, compared to the control knockdown, significantly inhibited cell proliferation (**Figure 6C**) and reduced cell numbers of SW620 cells (**Figures 6D, E**), which indicated that MAD2L1 might promote cell proliferation. To examine how MAD2L1affects cell growth, the cell cycle phase distribution and apoptosis were analyzed by flow cytometric analysis. Knockdown of MAD2L1 resulted in a decrease in the percentage of cells in the G1 and G2 phases and an increase in the percentage of cells in the S phase (**Figures 6F, G**), which indicated that MAD2L1 knockdown prevented cell passage from the S phase into the G2 phase. Therefore, MAD2L1 was shown to promote S/G2 phase transition. The apoptosis assay results indicated that the apoptotic cells significantly increased in SW480 cells with si-MAD2L1 transfection (**Figures 6H, I**). These data indicate that MAD2L1 knockdown could impair cell cycle progression and induce cell apoptosis.

DISCUSSION

Even with a gradual decline in the past few years, CRC remains the fourth leading cause of cancer-related death worldwide (Marmol et al., 2017). The occurrence and development of CRC is a dynamic process. At different stages of CRC, the expression levels of some molecules are different. (Moroishi et al., 2015) In this case, early screening and diagnosis are becoming increasingly difficult. Therefore, it is necessary to find accurate and meaningful CRC biomarkers. Our study systematically focused on expression profiles obtained from microarray studies of CRC. Our analysis included 74 CRC samples and 65 normal samples from the GSE117606 dataset of the GEO database. A total of 683 DEGs were identified, including 420 upregulated genes and 263 downregulated genes. To better explore these DEGs, we carried out GO function and KEGG pathway analysis of these DEGs.

GO analysis showed that the upregulated DEGs were particularly enriched in mitotic collagen catabolic process, extracellular matrix organization, proteinaceous extracellular matrix, extracellular region, extracellular matrix structural constituent, and extracellular matrix binding, while the downregulated DEGs were involved in bicarbonate transport, muscle contraction, extracellular exosome, extracellular space, chloride channel activity, and carbonate dehydratase activity. In addition, the KEGG pathways for the upregulated DEGs included the cell cycle, ECM-receptor interaction, and focal adhesion, while the pathways of the downregulated DEGs were mainly in mineral absorption, proximal tubule bicarbonate reclamation, and retinol metabolism.

A PPI is defined as the process by which two or more kinds of protein molecules form a protein complex by noncovalent bonding. PPI networks could provide a visible framework for a better understanding of the functional organization of the proteome (Liu et al., 2009). The enriched pathways of the top 3 modules showed that CRC was associated with the cell cycle-related pathway and the p53 signalling pathway.

Cell cycle-related genes that promote the proliferation of endothelial cells contribute to the progression of tumor growth and metastasis of CRC (Hong et al., 2009). *CDK1* encodes a serine/threonine kinase that controls the eukaryotic cell cycle by regulating mitotic onset, as well as the centrosome cycle (Santamaria et al., 2007). *CDK1* promotes cell proliferation *via* the phosphorylation and inhibition of the forkhead box O1 transcription factor (Liu et al., 2008). The alteration of CDK1 has been found in numerous cancer types, including breast cancer (Kim et al., 2008), esophageal adenocarcinoma (Hansel et al., 2005), hepatocellular carcinoma (Wu et al., 2019), pancreatic ductal adenocarcinoma (Piao et al., 2019), and oral squamous cell carcinoma (Chang et al., 2005). Iacopetta et al. revealed that p53 mutations that lose transactivation ability are more common in advanced CRC and associated with poor survival (Iacopetta et al., 2006). Slattery ML et al. suggested that the activation of p53 from cellular stress could target downstream genes that could in turn influence cell cycle arrest, apoptosis, and angiogenesis through mRNA:miRNA interactions (Slattery et al., 2018). In the p53 signaling pathway, the RRM2 gene was an oncogene that was overexpressed in colorectal cancer, with its elevated expression correlated with the invasion depth, poorly differentiated type, and tumor node metastasis stage (Lu et al., 2012).

Twenty DEGs with high connectivity were selected as hub genes for PPI network analysis. By analyzing the correlations and expression levels in GEPIA, we determined that the hub genes were obviously positively correlated and significantly overexpressed in CRC samples.

We searched the literature in PubMed for associations among the twenty hub genes in CRC. In Yanqi Gan et al.'s study, they revealed that expression of CCNA2 in CRC tissues is higher than that in normal tissues and that CCNA2 knockdown could significantly suppress CRC cell growth by impairing cell cycle progression and inducing cell apoptosis (Gan et al., 2018). *TOP2A* is a gene that involves copy number variations and chromosomal instability in many cancers (Simon et al., 2002; Bofin et al., 2003; Chen et al., 2015; Sonderstrup et al., 2015). In colorectal cancer, the protein expression level of *TOP2A* was related to aggressive tumor phenotypes and advanced tumor stages (Coss et al., 2009). In our research, we found that *TOP2A* expression was upregulated in colorectal cancer. The expression of PLK1 was correlated with tumor size, lymph node metastasis, depth of invasion, and TNM stage, consistent with the results from Takahashi et al. (Takahashi et al., 2003). Ding-pei Han et al.'s study revealed that PLK1 has additional functions and is involved in the proliferation, migration and invasion of colorectal cancer cells (Han et al., 2012). The spindle proteins AURKA, BUB1, and MAD2L1 are important components of the spindle assembly checkpoint (Xue et al., 2016), which has been frequently established as an important mechanism that drives aneuploidy and carcinogenesis in CRC (Chen et al., 1998; Burum-Auensen et al., 2007). Anke H, Sillars-Hardebol et al.'s study revealed TPX2 and AURKA as major players in this critical step in colorectal carcinogenesis (Sillars-Hardebol et al., 2012). RRM2 overexpression was significantly associated with invasion depth and differentiation, and clinical tissue specimens also showed that the expression levels of RRM2 may be associated with tumor stage, which was shown in Ai-Guo Lu et al.'s study (Lu et al., 2012). KIF11 is a mitotic kinesin and is required for the separation of duplicated centrosomes during spindle formation (Zhu et al., 2005). Imai T et al.'s results verified that knockdown of *KIF11* by siRNA inhibits sphere formation, indicating that *KIF11* is important in the activity of esophageal cancer and CRC (Imai et al., 2017). MELK was overexpressed and highly phosphorylated in colorectal adenocarcinomas, and its expression was significantly correlated with tumor stage and lymph node metastasis (Gong et al., 2018). NUSAP1 is a microtubule-binding protein that plays a vital role in the assembly of mitotic spindle (Song and Rape, 2010).

NUSAP1 gene silencing induced cell apoptosis and inhibited cell proliferation, cell migration, cell invasion, and EMT in colorectal cancer by inhibiting DNMT1 gene expression (Han et al., 2018). Human replication factor C (RFC) is a multimeric protein consisting of five distinct subunits that are highly conserved through evolution (Yao and O'Donnell, 2012). Jun Xiang et al.'s results revealed that the overexpression of *RFC4* commonly occurs in CRC and that a high expression level of RFC4 is associated with poor differentiation and late TNM stages in patients with CRC. Higher levels of RFC4 protein expression correlate with a worse overall survival in CRC (Xiang et al., 2014). Human pituitary tumor transforming gene-1 (PTTG1) is a novel oncogene. Ren Q et al.'s study preliminarily explored the effects of PTTG1 in colorectal cancer cell proliferation and metastasis and found that the downregulation of PTTG1 expression suppressed colorectal cancer cell proliferation, migration and invasion (Ren and Jin, 2017). Gali-Muhtasib H et al.'s study confirmed the *in vivo* existence of the CHEK1/p53 link in human colorectal cancer, showing that tumors lacking p53 had higher levels of CHEK1, which was accompanied by poorer apoptosis. CHEK1 overexpression was correlated with advanced tumor stages, proximal tumor localization, and worse prognosis (Gali-Muhtasib et al., 2008). Overexpression of *CEP55* activates p21 and enhances the cell cycle transition. In contrast, the knockdown of *CEP55* inhibits cell growth in gastric (Tao et al., 2014) and breast cancer (Wang et al., 2016). DTL is located at chromosomal region 1q32.1–32.2 and encodes a putative 730-amino-acid nuclear protein that contains six highly conserved WD40-repeat domains (Ueki et al., 2008). It has been reported that DTL plays an essential role in cell proliferation, cell cycle arrest and metastatic potential in hepatocellular carcinoma, breast cancer, gastric cancer and rhabdomyosarcoma (Pan et al., 2006; Ueki et al., 2008; Li et al., 2009; Missiaglia et al., 2009; Song et al., 2010). Baraniskin A et al.'s data identified miR-30a-5p as a tumor-suppressing miRNA in colon cancer cells, exerting its function *via* the modulation of DTL expression, which is frequently overexpressed in CRC (Baraniskin et al., 2012).

MAD2L1 is highly expressed in colon cancer according to biological information. Moreover, MAD2L1 has a high positive correlation, with a Pearson correlation coefficient of 0.88. Through bioinformatics analysis of GSE117606, we know that MAD2L1 is one of the 20 core genes, and that MAD2L1 plays a role in the occurrence and development of colon cancer by participating in the cell cycle pathway. In examining the expression level of MAD2L1, we found that MAD2L1 has a higher expression in the CRC clinical samples and cell lines. Afterward, by searching PubMed, we found that there were no relevant studies reporting that MAD2L1 is involved in the cell cycle pathway, so we chose MAD2L1 for the next cell experiments. We further confirmed that knockdown of MAD2L1 could significantly suppress CRC cell growth by impairing cell cycle progression and inducing cell apoptosis. MAD2L1 has the potential to be a new biomarker for diagnosis and therapy in CRC.

There is a limitation of this study that needs to be considered: the analysis of a single dataset from GEO will result in partial bias, and too few samples will not lead to new findings. However, the data set we selected contains a large number of samples, so this limitation can be compensated to a certain extent.

In summary, using the GSE117606 profile data set and multiple bioinformatics analyses, our present work identified twenty hub genes as DEGs. These DEGs are significantly enriched in several pathways that are mainly associated with the cell cycle, ECM-receptor interaction, and mineral absorption pathways in CRC, and they might play key roles in the development and progression of CRC. MAD2L1 shows higher expression levels in CRC, is involved in colon cancer cell growth and cell cycle progression, and could be used as a new biomarker since it has a significant meaning for clinical treatment.

CONCLUSION

In this study, using a GSE data set and multiple bioinformatics analyses, we identified twenty hub genes that were significantly enriched in the cell cycle, ECM–receptor interaction, and mineral absorption pathways in CRC. Moreover, the expression level of MAD2L1 was significantly increased in CRC, and knockdown of MAD2L1 suppressed colon cancer cell growth by impairing cell cycle and apoptosis progression. Our findings also establish that MAD2L1 could be a new biomarker for CRC diagnosis and guide combination therapy for CRC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117606).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of Renmin Hospital of Wuhan University Renmin Hospital of Wuhan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XD is responsible for the design of experiments, bioinformatic analysis, collection of samples and specific experimental operations. HD is responsible for data collation and statistical analysis. HL is responsible for providing experimental funds and technical guidance.

REFERENCES

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46, 175–185. doi: 10.1080/00031305.1992.10475879

Aran, V., Victorino, A. P., Thuler, L. C., and Ferreira, C. G. (2016). Colorectal cancer: epidemiology, disease mechanisms and interventions to reduce onset and mortality. *Clin. Colorectal Cancer* 15, 195–203. doi: 10.1016/j.clcc.2016.02.008

Baraniskin, A., Birkenkamp-Demtroder, K., Maghnouj, A., Zollner, H., Munding, J., Klein-Scory, S., et al. (2012). MiR-30a-5p suppresses tumor growth in colon carcinoma by targeting DTL. *Carcinogenesis* 33, 732–739. doi: 10.1093/carcin/bgs020

Bofin, A. M., Ytterhus, B., and Hagmar, B. M. (2003). TOP2A and HER-2 gene amplification in fine needle aspirates from breast carcinomas. *Cytopathology* 14, 314–319. doi: 10.1046/j.0956-5507.2003.00088.x

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Burum-Auensen, E., Deangelis, P. M., Schjolberg, A. R., Roislien, J., Andersen, S. N., Clausen, O. P., et al. (2007). Spindle proteins Aurora A and BUB1B, but not Mad2, are aberrantly expressed in dysplastic mucosa of patients with longstanding ulcerative colitis. *J. Clin. Pathol.* 60, 1403–1408. doi: 10.1136/jcp.2006.044305

Chang, J. T., Wang, H. M., Chang, K. W., Chen, W. H., and Wen, M. C. (2005). Identification of differentially expressed genes in oral squamous cell carcinoma (OSCC): overexpression of NPM, CDK1 and NDRG1 and underexpression of CHES1. *Int. J. Cancer* 114, 942–949. doi: 10.1002/ijc.20663

Chen, R. H., Shevchenko, A., Mann, M., and Murray, A. W. (1998). Spindle checkpoint protein Xmad1 recruits Xmad2 to unattached kinetochores. *J. Cell Biol.* 143, 283–295. doi: 10.1083/jcb.143.2.283

Chen, T., Sun, Y., Ji, P., Kopetz, S., and Zhang, W. (2015). Topoisomerase IIalpha in chromosome instability and personalized cancer therapy. *Oncogene* 34, 4019–4031. doi: 10.1038/onc.2014.332

Chen, L., Yuan, L., Wang, Y., Wang, G., Zhu, Y., Cao, R., et al. (2017). Co-expression network analysis identified FCER1G in association with progression and prognosis in human clear cell renal cell carcinoma. *Int. J. Biol. Sci.* 13, 1361–1372. doi: 10.7150/ijbs.21657

Coss, A., Tosetto, M., Fox, E. J., Sapetto-Rebow, B., Gorman, S., Kennedy, B. N., et al. (2009). Increased topoisomerase IIalpha expression in colorectal cancer is associated with advanced disease and chemotherapeutic resistance *via* inhibition of apoptosis. *Cancer Lett.* 276, 228–238. doi: 10.1016/j.canlet.2008.11.018

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Galili, T., O'Callaghan, A., Sidi, J., and Sievert, C. (2018). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* 34, 1600–1602. doi: 10.1093/bioinformatics/btx657

Gali-Muhtasib, H., Kuester, D., Mawrin, C., Bajbouj, K., Diestel, A., Ocker, M., et al. (2008). Thymoquinone triggers inactivation of the stress response pathway sensor CHEK1 and contributes to apoptosis in colorectal cancer cells. *Cancer Res.* 68, 5609–5618. doi: 10.1158/0008-5472.CAN-08-0884

Gan, Y., Li, Y., Li, T., Shu, G., and Yin, G. (2018). CCNA2 acts as a novel biomarker in regulating the growth and apoptosis of colorectal cancer. *Cancer Manag. Res.* 10, 5113–5124. doi: 10.2147/CMAR.S176833

Gaudet, P., Skunca, N., Hu, J. C., and Dessi, (2017). Primer on the gene ontology. *Methods Mol. Biol.* 1446, 25–37. doi: 10.1007/978-1-4939-3743-1_3

Gong, X., Chen, Z., Han, Q., Chen, C., Jing, L., Liu, Y., et al. (2018). Sanguinarine triggers intrinsic apoptosis to suppress colorectal cancer growth through disassociation between STRAP and MELK. *BMC Cancer* 18, 578. doi: 10.1186/s12885-018-4463-x

Han, D. P., Zhu, Q. L., Cui, J. T., Wang, P. X., Qu, S., Cao, Q. F., et al. (2012). Polo-like kinase 1 is overexpressed in colorectal cancer and participates in the migration and invasion of colorectal cancer cells. *Med. Sci. Monit.* 18, BR237–BR246. doi: 10.12659/MSM.882900

Han, G., Wei, Z., Cui, H., Zhang, W., Wei, X., Lu, Z., et al. (2018). NUSAP1 gene silencing inhibits cell proliferation, migration and invasion through inhibiting DNMT1 gene expression in human colorectal cancer. *Exp. Cell Res.* 367, 216–221. doi: 10.1016/j.yexcr.2018.03.039

Hansel, D. E., Dhara, S., Huang, R. C., Ashfaq, R., Deasel, M., Shimada, Y., et al. (2005). CDC2/CDK1 expression in esophageal adenocarcinoma and precursor lesions serves as a diagnostic and cancer progression marker and potential novel drug target. *Am. J. Surg. Pathol.* 29, 390–399. doi: 10.1097/00000478-200503000-00014

Hong, B. S., Cho, J. H., Kim, H., Choi, E. J., Rho, S., Kim, J., et al. (2009). Colorectal cancer cell-derived microvesicles are enriched in cell cycle-related mRNAs that promote proliferation of endothelial cells. *BMC Genomics* 10, 556. doi: 10.1186/1471-2164-10-556

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 8, R183. doi: 10.1186/gb-2007-8-9-r183

Iacopetta, B., Russo, A., Bazan, V., Dardanoni, G., Gebbia, N., Soussi, T., et al. (2006). Functional categories of TP53 mutation in colorectal cancer: results of an International Collaborative Study. *Ann. Oncol.* 17, 842–847. doi: 10.1093/annonc/mdl035

Imai, T., Oue, N., Sentani, K., Sakamoto, N., Uraoka, N., Egi, H., et al. (2017). KIF11 is required for spheroid formation by oesophageal and colorectal cancer cells. *Anticancer Res.* 37, 47–55. doi: 10.21873/anticanres.11287

Kanehisa,M.(2002)The KEGG database, in: Novartis Found Symp.247. pp. 91–101, discussion 101-3, 119-28, 244-52.

Kim, S. J., Nakayama, S., Miyoshi, Y., Taguchi, T., Tamaki, Y., Matsushima, T., et al. (2008). Determination of the specific activity of CDK1 and CDK2 as a novel prognostic indicator for early breast cancer. *Ann. Oncol.* 19, 68–72. doi: 10.1093/annonc/mdm358

Li, J., Ng, E. K., Ng, Y. P., Wong, C. Y., Yu, J., Jin, H., et al. (2009). Identification of retinoic acid-regulated nuclear matrix-associated protein as a novel regulator of gastric cancer. *Br. J. Cancer* 101, 691–698. doi: 10.1038/sj.bjc.6605202

Li, X. (1991). An aggregate function method for nonlinear programming. *Sci. In China* (A), 1467–1473.

Liu, P., Kao, T. P., and Huang, H. (2008). CDK1 promotes cell proliferation and survival *via* phosphorylation and inhibition of FOXO1 transcription factor. *Oncogene* 27, 4733–4744. doi: 10.1038/onc.2008.104

Liu, G., Wong, L., and Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics* 25, 1891–1897. doi: 10.1093/bioinformatics/btp311

Liu, M., Xu, Z., Du, Z., Wu, B., Jin, T., Xu, K., et al. (2017). The identification of key genes and pathways in glioma by bioinformatics analysis. *J. Immunol. Res.* 2017, 1278081. doi: 10.1155/2017/1278081

Lu, A. G., Feng, H., Wang, P. X., Han, D. P., Chen, X. H., Zheng, M. H., et al. (2012). Emerging roles of the ribonucleotide reductase M2 in colorectal cancer and ultraviolet-induced DNA damage repair. *World J. Gastroenterol.* 18, 4704–4713. doi: 10.3748/wjg.v18.i34.4704

Marmol, I., Sanchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., and Rodriguez Yoldi, M. J. (2017). Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int. J. Mol. Sci.* 18. doi: 10.3390/ijms18010197

Missiaglia, E., Selfe, J., Hamdi, M., Williamson, D., Schaaf, G., Fang, C., et al. (2009). Genomic imbalances in rhabdomyosarcoma cell lines affect expression of genes frequently altered in primary tumors: an approach to identify candidate genes involved in tumor development. *Genes Chromosomes Cancer* 48, 455–467. doi: 10.1002/gcc.20655

Moroishi, T., Hansen, C. G., and Guan, K. L. (2015). The emerging roles of YAP and TAZ in cancer. *Nat. Rev. Cancer* 15, 73–79. doi: 10.1038/nrc3876

Pan, H. W., Chou, H. Y., Liu, S. H., Peng, S. Y., Liu, C. L., Hsu, H. C., et al. (2006). Role of L2DTL, cell cycle-regulated nuclear and centrosome protein, in aggressive hepatocellular carcinoma. *Cell Cycle* 5, 2676–2687. doi: 10.4161/cc.5.22.3500

Piao, J., Zhu, L., Sun, J., Li, N., Dong, B., Yang, Y., et al. (2019). High expression of CDK1 and BUB1 predicts poor prognosis of pancreatic ductal adenocarcinoma. *Gene* 701, 15–22. doi: 10.1016/j.gene.2019.02.081

Ren, Q., and Jin, B. (2017). The clinical value and biological function of PTTG1 in colorectal cancer. *BioMed. Pharmacother.* 89, 108–115. doi: 10.1016/j.biopha.2017.01.115

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Santamaria, D., Barriere, C., Cerqueira, A., Hunt, S., Tardy, C., Newton, K., et al. (2007). Cdk1 is sufficient to drive the mammalian cell cycle. *Nature* 448, 811–815. doi: 10.1038/nature06046

Sillars-Hardebol, A. H., Carvalho, B., Tijssen, M., Belien, J. A., de Wit, M., Delis-van Diemen, P. M., et al. (2012). TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* 61, 1568–1575. doi: 10.1136/gutjnl-2011-301153

Simon, R., Atefy, R., Wagner, U., Forster, T., Fijan, A., Bruderer, J., et al. (2002). HER-2 and TOP2A gene amplification in urinary bladder cancer. *Verh. Dtsch. Ges. Pathol.* 86, 176–183.

Slattery, M. L., Mullany, L. E., Wolff, R. K., Sakoda, L. C., Samowitz, W. S., Herrick, J. S., et al. (2018). The p53-signaling pathway and colorectal cancer: interactions between downstream p53 target genes and miRNAs. *Genomics* 111 (4), 762–771. doi: 10.1016/j.ygeno.2018.05.006

Sonderstrup, I. M., Nygard, S. B., Poulsen, T. S., Linnemann, D., Stenvang, J., Nielsen, H. J., et al. (2015). Topoisomerase-1 and -2A gene copy numbers are elevated in mismatch repair-proficient colorectal cancers. *Mol. Oncol.* 9, 1207–1217. doi: 10.1016/j.molonc.2015.02.009

Song, L., Rape, M., Lu, A. G., Feng, H., Wang, P. X., Han, D. P., Chen, X. H., et al (2010). Regulated degradation of spindle assembly factors by the anaphase-promoting complex. *Mol. Cell* 38, 369–382. doi: 10.1016/j.molcel.2010.02.038

Song, B., Wang, Y., Titmus, M. A., et al. (2010). Molecular mechanism of chemoresistance by miR-215 in osteosarcoma and colon cancer cells. *Mol. Cancer* 9, 96. doi: 10.1186/1476-4598-9-96

Sun, C., Yuan, Q., Wu, D., Meng, X., and Wang, B. (2017). Identification of core genes and outcome in gastric cancer using bioinformatics analysis. *Oncotarget* 8, 70271–70280. doi: 10.18632/oncotarget.20082

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Takahashi, T., Sano, B., Nagata, T., Kato, H., Sugiyama, Y., Kunieda, K., et al. (2003). Polo-like kinase 1 (PLK1) is overexpressed in primary colorectal cancers. *Cancer Sci.* 94, 148–152. doi: 10.1111/j.1349-7006.2003.tb01411.x

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., Zhang, Z., et al. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

Tao, J., Zhi, X., Tian, Y., Li, Z., Zhu, Y., Wang, W., et al. (2014). CEP55 contributes to human gastric carcinoma by regulating cell proliferation. *Tumour Biol.* 35, 4389–4399. doi: 10.1007/s13277-013-1578-1

Ueki, T., Nishidate, T., Park, J. H., Lin, M. L., Shimo, A., Hirata, K., et al. (2008). Involvement of elevated expression of multiple cell-cycle regulator, DTL/RAMP (denticleless/RA-regulated nuclear matrix associated protein), in the growth of breast cancer cells. *Oncogene* 27, 5672–5683. doi: 10.1038/onc.2008.186

Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357. doi: 10.1126/science.aan2507

Wang, Y., Jin, T., Dai, X., and Xu, J. (2016). Lentivirus-mediated knockdown of CEP55 suppresses cell proliferation of breast cancer cells. *Biosci. Trends* 10, 67–73. doi: 10.5582/bst.2016.01010

Wu, M., Liu, Z., Li, X., Zhang, A., Lin, D., Li, N., et al. (2019). Analysis of potential key genes in very early hepatocellular carcinoma. *World J. Surg. Oncol.* 17, 77. doi: 10.1186/s12957-019-1616-6

Xiang, J., Fang, L., Luo, Y., Yang, Z., Liao, Y., Cui, J., et al. (2014). Levels of human replication factor C4, a clamp loader, correlate with tumor progression and predict the prognosis for colorectal cancer. *J. Transl. Med.* 12, 320. doi: 10.1186/s12967-014-0320-0

Xue, X., Ramakrishnan, S. K., Weisz, K., Triner, D., Xie, L., Attili, D., et al. (2016). Iron uptake *via* DMT1 integrates cell cycle with JAK-stat3 signaling to promote colorectal tumorigenesis. *Cell Metab.* 24, 447–461. doi: 10.1016/j.cmet.2016.07.015

Yao, N. Y., and O'Donnell, M. (2012). The RFC clamp loader: structure and function. *Subcell Biochem.* 62, 259–279. doi: 10.1007/978-94-007-4572-8_14

Zhu, C., Zhao, J., Bibikova, M., Leverson, J. D., Bossy-Wetzel, E., Fan, J. B., et al. (2005). Functional analysis of human microtubule-based motor proteins, the kinesins and dyneins, in mitosis/cytokinesis using RNA interference. *Mol. Biol. Cell* 16, 3187–3199. doi: 10.1091/mbc.e05-02-0167

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fgene.2019.01378

OSgbm: An Online Consensus Survival Analysis Web Server for Glioblastoma

Huan Dong[1†], Qiang Wang[1†], Ning Li[1], Jiajia Lv[1], Linna Ge[1], Mengsi Yang[1], Guosen Zhang[1], Yang An[1], Fengling Wang[1], Longxiang Xie[1], Yongqiang Li[1], Wan Zhu[2], Haiyu Zhang[3], Minghang Zhang[4] and Xiangqian Guo[1]*

[1] Department of Predictive Medicine, Institute of Biomedical Informatics, Cell Signal Transduction Laboratory, Bioinformatics Center, Henan Provincial Engineering Center for Tumor Molecular Medicine, School of Software, School of Basic Medical Sciences, Henan University, Kaifeng, China, [2] Department of Anesthesia, Stanford University School of Medicine, Stanford, CA, United States, [3] Department of Pathology, Stanford University School of Medicine, Stanford, CA, United States, [4] Nanjing Jiliang Biotechnology Co., Ltd., Nanjing, China

Glioblastoma (GBM) is the most common malignant tumor of the central nervous system. GBM causes poor clinical outcome and high mortality rate, mainly due to the lack of effective targeted therapy and prognostic biomarkers. Here, we developed a user-friendly Online Survival analysis web server for GlioBlastoMa, abbreviated OSgbm, to assess the prognostic value of candidate genes. Currently, OSgbm contains 684 samples with transcriptome profiles and clinical information from The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) and Chinese Glioma Genome Atlas (CGGA). The survival analysis results can be graphically presented by Kaplan-Meier (KM) plot with Hazard ratio (HR) and log-rank *p* value. As demonstration, the prognostic value of 51 previously reported survival associated biomarkers, such as *PROM1* (HR = 2.4120, $p = 0.0071$) and *CXCR4* (HR = 1.5578, $p < 0.001$), were confirmed in OSgbm. In summary, OSgbm allows users to evaluate and develop prognostic biomarkers of GBM. The web server of OSgbm is available at http://bioinfo.henu.edu.cn/GBM/GBMList.jsp.

Keywords: glioblastoma, survival analysis, prognostic biomarker, OSgbm, transcriptome profiles, clinical information

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Zhixiang Zuo,
Sun Yat-sen University, China
Xuexin Yu,
UT Southwestern Medical Center, United States

****Correspondence:***
Xiangqian Guo
xqguo@henu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *08 May 2019*
Accepted: *17 December 2019*
Published: *21 February 2020*

Citation:
Dong H, Wang Q, Li N, Lv J, Ge L, Yang M, Zhang G, An Y, Wang F, Xie L, Li Y, Zhu W, Zhang H, Zhang M and Guo X (2020) OSgbm: An Online Consensus Survival Analysis Web Server for Glioblastoma. Front. Genet. 10:1378. doi: 10.3389/fgene.2019.01378

INTRODUCTION

Glioblastoma (GBM) is the most common malignant tumor of the central nervous system (CNS) and causes a high mortality rate (Nikiforova and Hamilton, 2011; Stoyanov et al., 2018). Although many new therapies have improved the clinical outcome and more clinical trials have demonstrated the high efficacy in treating GBM, the survival rate of GBM patients is still low. GBM is a complex disease to tackle with a median survival period of approximately 14 months, and a 5-year survival rate of 5% (Stupp et al., 2005; Johnson and O'Neill, 2012; Polivka et al., 2017). Prognostic biomarkers have been showing great roles in cancer patient management and may guide targeted therapies. Therefore, it is greatly needed to investigate prognostic biomarkers in GBM.

Previous studies have reported some prognostic biomarkers in GBM, such as gene mutation of gene *IDH* and *PTEN*, and expression variation of gene *CD133* (Yang et al., 2016; Cai and Sughrue, 2017;

Nguyen et al., 2018). However, these biomarkers have not been translated to clinical applications due to the lack of independent validation. In addition, due to the molecular heterogeneity among GBMs and limited patient samples (Nathanson et al., 2014; Aldape et al., 2015; Brown et al., 2017), the prognostic behavior of a certain biomarker may be inconsistent or even contradictory between different reports. In other words, cross population validation in a larger patient cohort is critical for evaluating the prognostic biomarker.

In current work, we collected the gene expression profiles and clinical information of 684 GBM patients from seven independent cohorts obtained from TCGA, GEO and CGGA. We developed a user-friendly web server, OSgbm, to analyze the prognostic value of genes of interests. With this web server, it would facilitate researchers and clinicians to screen, develop and validate new prognostic biomarkers in GBM.

METHODS

Datasets Collection

GBM datasets are from three major data sources. First, level-3 gene expression profiling data (HiSeqV2) and clinical information of GBM samples were downloaded from TCGA on April 2018 (https://portal.gdc.cancer.gov/). Second, four cohorts (≥30 cases) with available gene expression profiles and clinical survival information were collected from GEO database (http://www.ncbi.nlm.nih.gov/geo/). Third, two GBM cohorts were gathered from CGGA (http://www.cgga.org.cn/). After an initial filtration and quality check (with available gene expression profiling data and clinical survival information), 153 samples from TCGA, 276 samples from GEO, and 255 samples from CGGA were included for the following database and web server construction. The histology of recurrent GBM (rGBM) were included in GSE7696 (10 samples), GSE42669 (11 samples), CGGAarray (9 samples) and CGGAseq (22 samples) datasets. Two CGGA datasets also included 20 samples of secondary GBM (sGBM).

System Implementation and Server Set-Up

OSgbm is a web-based tool which uses J2EE (Java 2 Platform Enterprise Edition) architecture as we previously described (Wang et al., 2019; Wang et al., 2019; Xie et al., 2019a; Zhang et al., 2019). The gene expression and clinical data were integrated in the background database, which was handled by a MySQL server. Dynamic web interfaces were written in HTML 5.0 and hosted by Tomcat on Windows Server. Using OSgbm requires a HTML 5.0-compliant browser with JavaScript enabled, but does not require any particular visual plug-in tool. Since the web server was designed for users with no specialized bioinformatics skills, we propose 'out-of-the-box' data. The input of OSgbm web server is official gene symbol. For the "Data Source: Combined" option, as all the datasets used in OSgbm already have been published, processed and normalized well, in order to avoid of the batch effect and platform biases among these datasets, we first stratify the patients into high- and low-expression group for the input gene in each dataset, and then merged relative patients from high- and low-expression group from each dataset into a combined high-expression group (Upper group in the Kaplan–Meier plot) and a combined low-expression group (Lower group in the Kaplan–Meier plot) for the analysis of Kaplan–Meier plot and log-rank test. The statistical analyses of input were performed with R package: KM curves with Hazard ratio (HR, 95% confidence interval) and log-rank p value were calculated by R package 'survival'. OSgbm is available at http://bioinfo.henu.edu.cn/GBM/GBMList.jsp.

Validation of Previously Reported Prognostic Biomarkers

A PubMed search was performed to identify previously reported GBM prognostic biomarkers, using keywords 'glioblastoma', 'survival' and 'biomarker'. Totally, 53 prognostic biomarkers were identified from 2013 publications. The flow chart of biomarker collection was showed in **Figure S1**. The prognostic values of these published biomarkers were analyzed in either a form of combined cohorts of all GBM patients or in a single cohort in our database.

RESULTS

The Clinical Characteristics of GBM Datasets Used in OSgbm

In OSgbm, we included a total of 684 unique GBM samples from seven datasets, including one TCGA cohort, four GEO cohorts and two CGGA cohorts. The survival information includes overall survival (OS), disease specific survival (DSS), disease free interval (DFI) and progression free interval (PFI) (Liu et al., 2018). The confounding clinical factors, such as age, grade, gender, histology and treatment regimens were included as well. Clinical characteristics of these datasets in the OSgbm were presented in **Table 1**. All of the 684 patients have OS data,

TABLE 1 | Clinical characteristics of each GBM dataset used in OSgbm.

Data Source	Sample Size (n)	Median Age (years)	Death (%)	OS Median (years)	Gender (male, %)	Grade (I/II/III/IV, %)	Survival Terms
TCGA	153	60	79.08	11.90	64.71	–	OS, DSS, DFI, PFI
GSE7696	80	52	81.25	15.58	73.75	–	OS
GSE4412	85	42	69.41	12.97	37.65	0/0/30.59/69.41	OS
GSE42669	57	51	80.70	14.93	52.63	–	OS
GSE30472	54	–	88.89	15.72	–	3.7/12.96/29.63/53.71	OS
CGGAseq	128	48	66.67	9.55	65.22	0/0/0.72/99.28	OS
CGGAarray	127	47	83.46	13.43	62.20	0/0/0/100	OS
Total	684	50	78.49	13.44	59.36	–	–

TABLE 2 | Validation of previously reported prognostic biomarkers in OSgbm.

Gene symbol	Validation results				Literature data				
	OS, HR (95% CI)	*p* Value	Cut Off	Osgbm	OS, HR (95% CI)	*p* Value	Sample (n)	Level	Reference
PROM1	2.412 (1.040–4.174)	0.007	Upper 25% vs Lower 25%	GSE7679	2.39 (1.77–3.23)	<0.001	656	mRNA	(Zhang et al., 2016)
SRGN	2.371 (1.256–4.477)	0.008	Upper 25% vs Lower 25%	CGGAseq	–	0.037	504	mRNA	(Roy et al., 2017)
EDNRB	2.272 (1.115–4.627)	0.024	Upper 25% vs Lower 75%	GSE30472	2.86 (1.12–7.34)	0.031	25	Protein	(Vasaikar et al., 2018)
PSMB4	2.074 (1.187–3.626)	0.010	Upper 25% vs Lower 25%	CGGAseq	–	<0.001	77	Protein	(Cheng et al., 2018)
WNT6	2.035 (1.098–3.770)	0.024	Upper 25% vs Lower 25%	CGGAseq	–	0.004	16	Protein	(Gonçalves et al., 2018)
DPYSL5	2.023 (1.160–3.527)	0.013	Upper 25% vs Lower 25%	CGGAarray	–	0.026	183	Protein	(Moutal et al., 2015)
IL17A	2.009 (1.107–3.646)	0.022	Upper 50% vs Lower 50%	GSE30472	–	0.007	41	Protein	(Cui et al., 2013)
TLR9	1.976 (1.089–3.588)	0.025	Upper 25% vs Lower 25%	CGGAseq	–	0.020	46	Protein	(Mu et al., 2017)
ACKR3	1.974 (1.040–3.747)	0.038	Upper 30% vs Lower 30%	GSE7679	1.56 (1.04–2.51)	0.03	146	Protein	(Deng et al., 2017)
H19	1.864 (1.309–2.653)	<0.001	Upper 25% vs Lower 25%	Combined	–	0.034	–	mRNA	(Wu et al., 2017)
EGFR	1.845 (1.077–3.160)	0.026	Upper 25% vs Lower 75%	GSE7696	–	<0.001	196	Protein	(Heimberger et al., 2005)
NUSAP1	1.748 (1.006–3.040	0.048	Upper 25% vs Lower 25%	CGGAarray	0.65 (0.49–0.86)*	0.003	518	mRNA	(Qian et al., 2018)
CHAF1B	1.707 (1.323–2.203)	<0.001	Upper 30% vs Lower 30%	Combined	–	0.004	96	Protein	(De Tayrac et al., 2013)
TAGLN2	1.665 (1.282–2.161)	<0.001	Upper 25% vs Lower 25%	Combined	–	<0.05	667	mRNA	(Han et al., 2017)
BIRC1	1.658 (1.266–2.172)	<0.001	Upper 25% vs Lower 25%	Combined	–	0.0003	66	Protein	(Shirai et al., 2009)
MGMT	1.633 (1.260–2.115)	<0.001	Upper 25% vs Lower 25%	Combined	1.50	0.01	157	Protein	(Dahlrot et al., 2018)
CD70	1.561 (1.180–2.065)	0.002	Upper 25% vs Lower 25%	Combined	1.6 (0.98–2.51)	0.046	107	mRNA	(Ge et al., 2017)
CXCR4	1.558 (1.207–2.010)	<0.001	Upper 25% vs Lower 25%	Combined	–	<0.05	156	mRNA	(Ma et al., 2017)
CA9	1.556 (1.202–2.015)	<0.001	Upper 25% vs Lower 75%	Combined	–	0.004	66	Protein	(Cetin et al., 2018)
PDCD1	1.508 (1.171–1.942)	0.002	Upper 30% vs Lower 30%	Combined	–	0.028	149	mRNA	(Nduom et al., 2016)
IDH1	1.490 (1.013–2.192)	0.043	Upper 50% vs Lower 50%	CGGAarray	–	0.045	163	Protein	(Chaurasia et al., 2016)
IGFBP2	1.467 (1.132–1.902)	0.004	Upper 25% vs Lower 25%	Combined	1.04 (1.02–1.05)	0.001	83	Plasma	(Han et al., 2014)
PBK	1.456 (1.131–1.875)	0.004	Upper 25% vs Lower 25%	Combined	–	0.007	32	Protein	(Hayashi et al., 2018)
EFEMP2	1.446 (1.117–1.871)	0.005	Upper 25% vs Lower 25%	Combined	–	<0.01	77	mRNA	(Li et al., 2017)
MET	1.434 (1.130–1.820)	0.003	Upper 30% vs Lower 30%	Combined	1.7 (1.1–2.2)	<0.05	69	Protein	(Olmez et al., 2014)
CHI3L1	1.438 (1.104–1.872)	0.007	Upper 25% vs Lower 25%	GSE30472	–	<0.01	98	mRNA	(Steponaitis et al., 2016)
TRAF2	1.443 (1.118–1.863)	0.005	Upper 25% vs Lower 25%	Combined	–	0.03	105	mRNA	(Zhang et al., 2017)
HMGB2	1.391 (1.099–1.759)	0.006	Upper 30% vs Lower 30%	Combined	3.35 (1.25–9.02)	0.017	51	Protein	(Wu et al., 2013)
MCM6	1.387 (1.132–1.699)	0.002	Upper 25% vs Lower 75%	Combined	1.19	0.006	325	mRNA	(Cai et al., 2018)
CD44	1.386 (1.073–1.790)	0.012	Upper 25% vs Lower 25%	Combined	–	<0.001	28	Protein	(Steponaitis et al., 2016)

(Continued)

TABLE 2 | Continued

Gene symbol	Validation results				Literature data				
	OS, HR (95% CI)	*p* Value	Cut Off	Osgbm	OS, HR (95% CI)	*p* Value	Sample (n)	Level	Reference
TIMP1	1.342 (1.025–1.758)	0.033	Upper 25% vs Lower 25%	Combined	3.2 (1.5–6.7)	0.004	112	Protein	(Aaberg-Jessen et al., 2009)
CD151	1.336 (1.023–1.746)	0.034	Upper 25% vs Lower 25%	Combined	5.064 (1.427–17.969)	0.012	211	Protein	(Lee et al., 2013)
TWIST1	1.312 (1.013–1.699)	0.039	Upper 25% vs Lower 25%	Combined	5.745 (1.331–1.89)	0.017	86	Protein	(Wang et al., 2013)
CCT6A	1.316 (1.045–1.655)	0.019	Upper 30% vs Lower 30%	Combined	3.21 (2.85–3.65)	0.006	497	Protein	(Hallal et al., 2019)
APC	1.308 (1.093–1.566)	0.004	Upper 50% vs Lower 50%	Combined	–	<0.001	83	Protein	(Rosati et al., 2013)
CD247	1.292 (1.022–1.633)	0.032	Upper 30% vs Lower 30%	Combined	1.54 (1.05–2.28)	0.023	149	mRNA	(Nduom et al., 2016)
CXCR3	1.272 (1.027–1.575)	0.028	Upper 25% vs Lower 75%	Combined	1.56 (1.04–2.51)	0.03	146	Protein	(Pu et al., 2011)
TCTN1	1.223 (1.011–1.493)	0.039	Upper 30% vs Lower 70%	Combined	1.32 (1.08–1.61)	0.006	518	mRNA	(Meng et al., 2014)
BICD1	0.794 (0.644–0.978)[#]	0.030	Lower 25% vs Upper 75%	Combined	1.577 (1.299–1.914)	<0.001	523	mRNA	(Huang et al., 2017)
IFIT1	0.770 (0.609–0.973)	0.029	Upper 30% vs Lower 30%	Combined	0.22 (0.10–0.52)	0.001	70	mRNA	(Zhang et al., 2016)
BRMS1L	0.753 (0.587–0.966)	0.026	Upper 25% vs Lower 75%	Combined	–	<0.05	60	mRNA	(Lv et al., 2018)
IGF1R	0.745 (0.588–0.944)	0.015	Upper 30% vs Lower 30%	Combined	1.65 (1.10–2.47)	0.016	167	Protein	(Maris et al., 2015)
GANO1	0.748 (0.585–0.957)	0.021	Upper 30% vs Lower 30%	Combined	–	0.009	178	Protein	(Zupancic et al., 2014)
PTEN	0.729 (0.567–0.938)	0.014	Upper 25% vs Lower 25%	Combined	3.3 (1.6–4.3)*	0.0003	61	mRNA	(Sano et al., 1999)
SEMA6A	0.694 (0.556–0.867)	0.001	Upper 25% vs Lower 75%	Combined	1.71 (1.01–2.65)*	0.012	200	Protein	(Zhao et al., 2015)
PHF3	0.683 (0.529–0.883)	0.004	Upper 25% vs Lower 25%	Combined	0.44 (0.26–0.77)	0.0031	35	Protein	(Yan et al., 2015)
PPARα	0.644 (0.503–0.825)	<0.001	Upper 30% vs Lower 30%	Combined	1.31 (1.05–1.63)*	0.016	473	mRNA	(Haynes et al., 2017)
PCBP2	0.632 (0.417–0.957)	0.031	Upper 25% vs Lower 75%	TCGA	–	<0.001	130	mRNA	(Luo and Zhuang, 2017)
LAPTM4B	0.626 (0.433–0.894)	0.010	Upper 50% vs Lower 50%	TCGA	–	<0.001	39	Protein	(Dong et al., 2017)
ANXA7	0.619 (0.475–0.806)	<0.001	Upper 25% vs Lower 25%	Combined	–	<0.001	99	Protein	(Hung and Howng, 2003)
PHF20	0.557 (0.319–0.972)	0.040	Upper 50% vs Lower 50%	CGGAarray	0.5 (0.29–0.86)	0.012	62	Protein	(Yan et al., 2015)
TES	0.407 (0.173–0.958)	0.040	Upper 30% vs Lower 30%	GSE42669	–	<0.05	37	Protein	(Bai et al., 2014)
LGALS1	0.368 (0.157–0.863)	0.022	Upper 25% vs Lower 25%	GSE42669	–	0.009	45	Protein	(Chou et al., 2018)

*: *The lower gene expression compared with higher gene expression in the literature data.*
#: *The lower gene expression compared with higher gene expression in the OSgbm data.*

and the median OS time was 13.44 months, while 153 GBM patients from TCGA cohort have four above mentioned survival terms (OS, DSS, DFI and PFI). The median age of all the patients is 50 years. The death rate is 78.49%. A large proportion of the patients are in grade IV, especially in the two CGGA datasets (99.28% and 100%, respectively).

Set-Up of OSgbm Web Server

The main function of OSgbm web server is to evaluate and determine the prognostic value of the quested genes. The users start by typing the gene symbol and choosing one dataset of interest or the combined dataset with pooling all the datasets together. To measure the association between a quested gene and survival, GBM samples are categorized according to the median (or other appropriate cutoff value, such as Trichotomy, Quartile) of the selected gene, and KM analysis is used to compare the outcomes between groups (Xie et al., 2019b). The user could limit the analysis in a subgroup of the patients by setting the age range, grade, gender and so on. Once the gene symbol is input and clinical characters are chosen, OS, DSS, DFI or PFI of each

stratified group can be measured and analysis results will be available on the output web page. The prognostic value of each given gene is determined by HR (95% CI) and log-rank *p* value.

Validation of Previously Reported GBM Prognostic Biomarkers

To determine the performance of this online tool, 53 previously published GBM prognostic factors collected as the procedure shown in **Figure S1** and then they were evaluated in OSgbm (**Table 2**, **Figure 1**) (Sano et al., 1999; Hung and Howng, 2003; Heimberger et al., 2005; Aaberg-Jessen et al., 2009; Shirai et al., 2009; Pu et al., 2011; Cui et al., 2013; De Tayrac et al., 2013; Lee et al., 2013; Rosati et al., 2013; Wang et al., 2013; Wu et al., 2013; Bai et al., 2014; Han et al., 2014; Meng et al., 2014; Olmez et al., 2014; Zupancic et al., 2014; Maris et al., 2015; Moutal et al., 2015; Yan et al., 2015; Zhao et al., 2015; Chaurasia et al., 2016; Nduom et al., 2016; Steponaitis et al., 2016; Steponaitis et al., 2016; Zhang et al., 2016a; Zhang et al., 2016b; Deng et al., 2017; Dong et al., 2017; Ge et al., 2017; Han et al., 2017; Haynes et al., 2017; Huang et al., 2017; Li et al., 2017; Luo and Zhuang, 2017; Ma et al., 2017; Mu et al., 2017; Roy et al., 2017; Wu et al., 2017; Zhang et al., 2017; Cai et al., 2018; Cetin et al., 2018; Cheng et al., 2018; Chou et al., 2018; Dahlrot et al., 2018; Gonçalves et al., 2018; Hayashi et al., 2018; Lv et al., 2018; Qian et al., 2018; Vasaikar et al., 2018; Hallal et al., 2019). OS was selected as the survival term. Among these prognostic genes, 51 of them showed significant prognostic ability in a large-scale combined cohort (33 genes) or in single cohort (18 genes), which were consistent with the prognostic value reported in the literature. The remaining two genes (*IGF1R* and *PCBP2*) display significant prognostic values in OSgbm, but is contradictory to what was reported in the literatures. Both of them were shown as favorable prognostic biomarkers in OSgbm

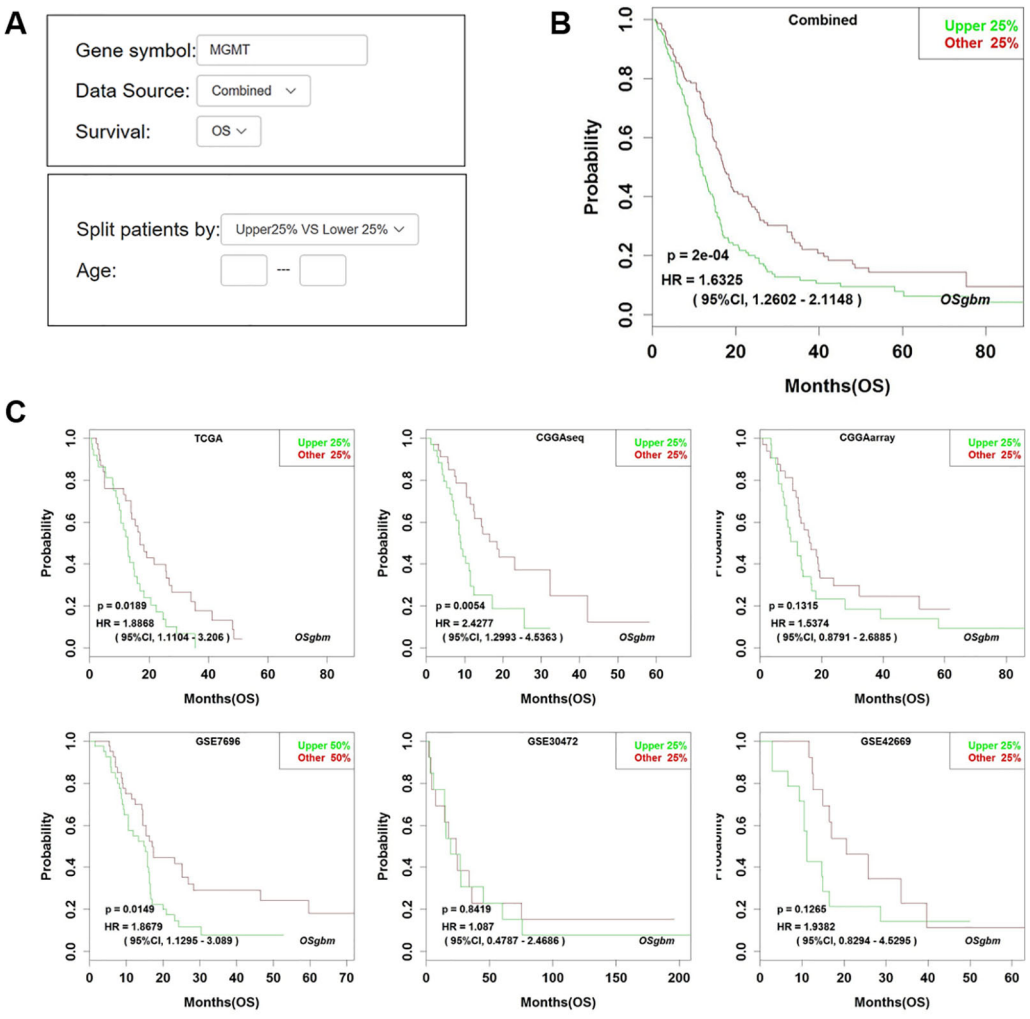


FIGURE 1 | Analysis of the prognostic value of *MGMT* in OSgbm. **(A)** The options of input parameters used in the prognostic analysis of *MGMT* in OSgbm. **(B)** The output web page of prognosis analysis of *MGMT* using a combined cohort with pooling all datasets together in OSgbm. **(C)** The OSgbm output of gene *MGMT* in single cohort.

but were reported to be unfavorable GBM prognostic biomarkers in previous reports (**Table 2**) (Maris et al., 2015; Luo and Zhuang, 2017).

DISCUSSION

The development of prognostic biomarkers is important for guiding the treatments especially for therapy-resistant GBM patients. In our work, we developed a new web server, OSgbm, to help researchers to evaluate the prognostic value of a given gene for GBM patients. OSgbm is easy to use and requires no special skills (such as bioinformatics training). With filtering by one or several clinical confounding factors provided in OSgbm, users can also evaluate the prognostic value of their interested genes according to their special needs. The function and performance tests of OSgbm web server showed that 96% (51 out of 53) of previously reported prognostic biomarkers could be confirmed in OSgbm, which indicates that these biomarkers validated in independent cohorts have the potency of translating to clinical applications, and also indicates the well performance of OSgbm. Nevertheless, there are two genes including *IGF1R* and *PCBP2* which showed different prognostic values to the literatures, the discrepancy of prognostic performance of *IGF1R* and *PCBP2* between OSgbm and literatures may be caused by race, different cohort size, or analysis level and methods (mRNA vs. protein, gene microarray vs. immunohistochemistry) (Maris et al., 2015; Luo and Zhuang, 2017). For example, the race reported in literatures for *PCBP2* is Asian, while that in validated cohort of OSgbm is mostly White. The mRNA level was analyzed in OSgbm for *IGF1R*, while IGF1R was determined by immunohistochemistry in literature. In addition, the race analyzed in OSgbm for *IGF1R* is Asian (Korea for GSE42669 and Chinese for CGGA), while the race reported in literature for IGF1R is European. As a result, it will be necessary to validate the prognostic performance of *IGF1R* and *PCBP2* in a larger independent cohort of glioblastoma.

In conclusion, OSgbm is a user-friendly web server to help researchers and clinicians to identify suitable prognostic biomarkers in GBM. Furthermore, we will keep update the database of OSgbm to collect more and more GBM datasets when new GBM dataset is available, and will implement the multivariate cox proportional hazards model into OSgbm for the purpose of adjustment for the confounding clinical factors, and we also encourage users to contact us to upload their own data into OSgbm.

DATA AVAILABILITY STATEMENT

All datasets for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

XG conceived and directed the project. HD and QW collected data and developed the web server. HD, NL, JL, LG, MY, GZ, YA, FW, LX, and YL performed data analysis. WZ, HZ, and MZ contributed to data analysis and paper writing. XG and HD wrote the manuscript with the assistance and approval of all authors.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 81602362), the program for Science and Technology Development in Henan Province (No. 162102310391), the supporting grants of Henan University (No. 2015YBZR048; No. B2015151), the program for Innovative Talents of Science and Technology in Henan Province (No. 18HASTIT048), and Yellow River Scholar Program (No. H2016012), Kaifeng Science and Technology Major Project (No. 18ZD008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01378/full#supplementary-material

REFERENCES

Aaberg-Jessen, C., Christensen, K., Offenberg, H., Bartels, A., Dreehsen, T., Hansen, S., et al. (2009). Low expression of tissue inhibitor of metalloproteinases-1 (TIMP-1) in glioblastoma predicts longer patient survival. *J. Neurooncol.* 95, 117–128. doi: 10.1007/s11060-009-9910-8

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., and von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms and markers. *Acta Neuropathol.* 129, 829–848. doi: 10.1007/s00401-015-1432-1

Bai, Y., Zhang, Q., and Wang, X. (2014). Downregulation of TES by hypermethylation in glioblastoma reduces cell apoptosis and predicts poor clinical outcome. *Eur. J. Med. Res.* 19, 66. doi: 10.1186/s40001-014-0066-4

Brown, D. V., Filiz, G., Daniel, P. M., Hollande, F., Dworkin, S., Amiridis, S., et al. (2017). Expression of CD133 and CD44 in glioblastoma stem cells correlates with cell proliferation, phenotype stability and intra-tumor heterogeneity. *PloS One* 12, e0172791. doi: 10.1371/journal.pone.0172791

Cai, X., and Sughrue, M. E. (2017). Glioblastoma: new therapeutic strategies to address cellular and genomic complexity. *Oncotarget* 20, 9540–9554. doi: 10.18632/oncotarget.23476

Cai, H., Cheng, Z., Zhang, H., Wang, P., Zhang, Y., Hao, J., et al. (2018). Overexpression of MCM6 predicts poor survival in patients with glioma. *Hum. Pathol.* 78, 182–187. doi: 10.1016/j.humpath.2018.04.024

Cetin, B., Gonul, I. I., Gumusay, O., Bilgetekin, I., Algin, E., Ozet, A., et al. (2018). Carbonic anhydrase IX is a prognostic biomarker in glioblastoma multiforme. *Neuropathology* 38, 457–462. doi: 10.1111/neup.12485.

Chaurasia, A., Park, S. H., Seo, J. W., and Park, C. K. (2016). Immunohistochemical analysis of ATRX, IDH1 and p53 in glioblastoma and their correlations with patient survival. *J. Korean Med. Sci.* 31, 1208–1214. doi: 10.3346/jkms.2016.31.8.1208

Cheng, Y., Tsai, W., Sung, Y., Chang, H., and Chen, Y. (2018). Interference with PSMB4 expression exerts an anti-tumor effect by decreasing the invasion and proliferation of human glioblastoma cells. *Cell Physiol. Biochem.* 45, 819–831. doi: 10.1159/000487174

Chou, S., Yen, S., Huang, C., and Huang, E. (2018). Galectin-1 is a poor prognostic factor in patients with glioblastoma multiforme after radiotherapy. *BMC Cancer* 18, 105. doi: 10.1186/s12885-018-4025-2

Cui, X., Xu, Z., Zhao, Z., Sui, D., Ren, X., Huang, Q., et al. (2013). Analysis of CD137L and IL-17 expression in tumor tissue as prognostic indicators for gliblastoma. *Int. J. Biol. Sci.* 9, 134–141. doi: 10.7150/ijbs.4891

Dahlrot, R. H., Dowsett, J., Fosmark, S., Malmström, A., Henriksson, R., Boldt, H., et al. (2018). Prognostic value of O-6-methylguanine-DNA methyltransferase (MGMT) protein expression in glioblastoma excluding nontumour cells from the analysis. *Neuropathol. Appl. Neurobiol.* 44, 172–184. doi: 10.1111/nan.12415

De Tayrac, M., Saikali, S., Aubry, M., Bellaud, P., Boniface, R., Quillien, V., et al. (2013). Prognostic significance of EDN/RB, HJURP, p60/CAF-1 and PDLI4, four new markers in high-grade gliomas. *PloS One* 8, e73332. doi: 10.1371/journal.pone.0073332

Deng, L., Zheng, W., Dong, X., Liu, J., Zhu, C., Lu, D., et al. (2017). Chemokine receptor CXCR7 is an independent prognostic biomarker in glioblastoma. *Cancer Biomark.* 20, 1–6. doi: 10.3233/CBM-151430

Dong, X., Tamura, K., Kobayashi, D., Ando, N., Sumita, K., and Maehara, T. (2017). LAPTM4B-35 is a novel prognostic factor for glioblastoma. *J. Neurooncol.* 132, 295–303. doi: 10.1007/s11060-017-2369-0

Ge, H., Mu, L., Jin, L., Yang, C., Chang, Y., Long, Y., et al. (2017). Tumor associated CD70 expression is involved in promoting tumor migration and macrophage infiltration in GBM. *Int. J. Cancer* 141, 1434–1444. doi: 10.1002/ijc.30830

Gonçalves, C. S., Vieira de Castro, J., Pojo, M., Martins, E. P., Queirós, S., Chautard, E., et al. (2018). WNT6 is a novel oncogenic prognostic biomarker in human glioblastoma. *Theranostics* 8, 4805–4823. doi: 10.7150/thno.25025

Hallal, S., Russell, B. P., Wei, H., Lee, M. Y. T., Toon, C. W., Sy, J., et al. (2019). Extracellular Vesicles from neurosurgical aspirates identifies chaperonin containing TCP1 subunit 6A as a potential glioblastoma biomarker with prognostic significance. *Proteomics* 19, e1800157. doi: 10.1002/pmic.201800157

Han, S., Meng, L., Han, S., Wang, Y., and Wu, A. (2014). Plasma IGFBP-2 levels after postoperative combined radiotherapy and chemotherapy predict prognosis in elderly glioblastoma patients. *PloS One* 9, e93791. doi: 10.1371/journal.pone.0093791

Han, M., Xu, R., Xu, Y., Zhang, X., Ni, S., Huang, B., et al. (2017). TAGLN2 is a candidate prognostic biomarker promoting tumorigenesis in human gliomas. *J. Exp. Clin. Cancer Res.* 36, 155. doi: 10.1186/s13046-017-0619-9

Hayashi, T., Hayakawa, Y., Koh, M., Tomita, T., Nagai, S., Kashiwazaki, D., et al. (2018). Impact of a novel biomarker, T-LAK cell-originating protein kinase (TOPK) expression on outcome in malignant glioma. *Neuropathology* 38, 144–153. doi: 10.1111/neup.12446

Haynes, H. R., White, P., Hares, K. M., Redondo, J., Kemp, K. C., Singleton, W. G. B., et al. (2017). The transcription factor PPARα is overexpressed and is associated with a favourable prognosis in IDH-wildtype primary glioblastoma. *Histopathology* 70, 1030–1043. doi: 10.1111/his.13142

Heimberger, A. B., Hlatky, R., Suki, D., Yang, D., Weinberg, J., Gilbert, M., et al. (2005). Prognostic effect of epidermal growth factor receptor and EGFRvIII in glioblastoma multiforme patients. *Clin. Cancer Res.* 11, 1462–1466. doi: 10.1158/1078-0432.CCR-04-1737

Huang, S. P., Chang, Y. C., Low, Q. H., Wu, A. T. H., Chen, C. L., Lin, Y. F., et al. (2017). BICD1 expression, as a potential biomarker for prognosis and predicting response to therapy in patients with glioblastomas. *Oncotarget* 8, 113766–113791. doi: 10.18632/oncotarget.22667

Hung, K. S., and Howng, S. L. (2003). Prognostic significance of annexin VII expression in glioblastomas multiforme in humans. *J. Neurosurg.* 99, 886–892. doi: 10.3171/jns.2003.99.5.0886

Johnson, D. R., and O'Neill, B. P. (2012). Glioblastoma survival in the United States before and during the temozolomide era. *J. Neurooncol.* 107, 359–364. doi: 10.1002/cncr.26494

Lee, D., Suh, Y. L., Park, T. I., Do, I. G., Seol, H. J., Nam, D. H., et al. (2013). Prognostic significance of tetraspanin CD151 in newly diagnosed glioblastomas. *J. Surg. Oncol.* 107, 646–652. doi: 10.1002/jso.23249

Li, F., Li, Y., Zhang, K., Li, Y., He, P., Liu, Y., et al. (2017). FBLN4 as candidate gene associated with long-term and short-term survival with primary glioblastoma. *Onco. Targets Ther.* 10, 387–395. doi: 10.2147/OTT.S117165

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416. doi: 10.1016/j.cell.2018.02.052

Luo, K., and Zhuang, K. (2017). High expression of PCBP2 is associated with progression and poor prognosis in patients with glioblastoma. *BioMed. Pharmacother.* 94, 659–665. doi: 10.1016/j.biopha.2017.07.103

Lv, J., Yang, H., Wang, X., He, R., Ding, L., and Sun, X. (2018). Decreased BRMS1L expression is correlated with glioma grade and predicts poor survival in glioblastoma *via an* invasive phenotype. *Cancer Biomark.* 22, 311–316. doi: 10.3233/CBM-171019

Ma, X., Shang, F., Zhu, W., and Lin, Q. (2017). CXCR4 expression varies significantly among different subtypes of glioblastoma multiforme (GBM) and its low expression or hypermethylation might predict favorable overall survival. *Expert Rev. Neurother* 7, 941–946. doi: 10.1080/14737175.2017.1351299

Maris, C., D'Haene, N., Trépant, A. L., Le Mercier, M., Sauvage, S., and Allard, J. (2015). IGF-IR: a new prognostic biomarker for human glioblastoma. *Br. J. Cancer* 113, 729–737. doi: 10.1038/bjc.2015.242

Meng, D., Chen, Y., Zhao, Y., Wang, J., Yun, D., Yang, S., et al. (2014). Expression and prognostic significance of TCTN1 in human glioblastoma. *J. Transl. Med.* 12, 288. doi: 10.1186/s12967-014-0288-9

Moutal, A., Honnorat, J., Massoma, P., Désormeaux, P., Bertrand, C., Malleval, C., et al. (2015). CRMP5 controls glioblastoma cell proliferation and survival through notch-dependent signaling. *Cancer Res.* 75, 3519–3528. doi: 10.1158/0008-5472.CAN-14-0631

Mu, L., Wang, Y., Wang, Y., Zhang, H., Shang, D., Tan, F., et al. (2017). Tumor location and survival outcomes in adult patients with supratentorial glioblastoma by levels of toll-like receptor 9 expression. *World Neurosurg.* 97, 279–283. doi: 10.1016/j.wneu.2016.10.015

Nathanson, D. A., Gini, B., Mottahedeh, J., Visnyei, K., Koga, T., Gomez, G., et al. (2014). Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* 343, 72–76. doi: 10.1126/science.1241328

Nduom, E. K., Wei, J., Yaghi, N. K., Huang, N., Kong, L. Y., Gabrusiewicz, K., et al. (2016). PD-L1 expression and prognostic impact in glioblastoma. *Neuro Oncol.* 18, 195–205. doi: 10.1093/neuonc/nov172

Nguyen, H. S., Shabani, S., Awad, A. J., Kaushal, M., and Doan, N. (2018). Molecular markers of therapy-resistant glioblastoma and potential strategy to combat resistance. *Int. J. Mol. Sci.* 19. doi: 10.3390/ijms19061765

Nikiforova, M. N., and Hamilton, R. L. (2011). Molecular diagnostics of gliomas. *Arch. Pathol. Lab. Med.* 135, 558–568. doi: 10.1043/2010-0649-RAIR.1

Olmez, O. F., Cubukcu, E., Evrensel, T., Kurt, M., Avci, N., Tolunay, S., et al. (2014). The immunohistochemical expression of c-Met is an independent predictor of survival in patients with glioblastoma multiforme. *Clin. Transl. Oncol.* 16, 173–177. doi: 10.1007/s12094-013-1059-4

Polivka, JJr, Polivka, J., Holubec, L., Kubikova, T., Priban, V., Hes, O., et al. (2017). Advances in experimental targeted therapy and immunotherapy for patients with glioblastoma multiforme. *Anticancer Res.* 37, 21–33. doi: 10.21873/anticanres.11285

Pu, Y., Li, S., Zhang, C., Bao, Z., Yang, Z., and Sun, L. (2011). High expression of CXCR3 is an independent prognostic factor in glioblastoma patients that promotes an invasive phenotype. *J. Neurooncol.* 122, 43–51. doi: 10.1007/s11060-014-1692-y

Qian, Z., Li, Y., Ma, J., Xue, Y., Xi, Y., et al. (2018). Prognostic value of NUSAP1 in progression and expansion of glioblastoma multiforme. *J. Neurooncol.* 140, 199–208. doi: 10.1007/s11060-018-2942-1

Rosati, A., Poliani, P. L., Todeschini, A., Cominelli, M., Medicina, D., Cenzato, M., et al. (2013). Glutamine synthetase expression as a valuable marker of epilepsy and longer survival in newly diagnosed glioblastoma multiforme. *Neuro Oncol.* 15, 618–625. doi: 10.1093/neuonc/nos338.

Roy, A., Attarha, S., Weishaupt, H., Edqvist, P. H., Swartling, F. J., Bergqvist, M., et al. (2017). Serglycin as a potential biomarker for glioma: association of serglycin expression, extent of mast cell recruitment and glioblastoma progression. *Oncotarget* 11, 24815–24827. doi: 10.18632/oncotarget.15820

Sano, T., Lin, H., Chen, X., Langford, L. A., Koul, D., Bondy, M. L., et al. (1999). Differential expression of MMAC/PTEN in glioblastoma multiforme: relationship to localization and prognosis. *Cancer Res.* 59, 1820–1824.

Shirai, K., Suzuki, Y., Oka, K., Noda, S. E., Katoh, H., Suzuki, Y., et al. (2009). Nuclear survivin expression predicts poorer prognosis in glioblastoma. *J. Neurooncol.* 91, 353–358. doi: 10.1007/s11060-008-9720-4

Steponaitis, G., Skiriutė, D., Kazlauskas, A., Golubickaitė, I., Stakaitis, R., Tamašauskas, A., et al. (2016). High CHI3L1 expression is associated with glioma patient survival. *Diagn. Pathol.* 11, 42. doi: 10.1186/s13000-016-0492-4

Stoyanov, G. S., Dzhenkov, D., Ghenev, P., Iliev, B., Enchev, Y., and Tonchev, A. B. (2018). Cell biology of glioblastoma multiforme: from basic science to diagnosis and treatment. *Med. Oncol.* 35, 27. doi: 10.1007/s12032-018-1083-x

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl. J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Vasaikar, S., Tsipras, G., Landázuri, N., Costa, H., Wilhelmi, V., Scicluna, P., et al. (2018). Overexpression of endothelin B receptor in glioblastoma: a prognostic marker and therapeutic target? *BMC Cancer* 18, 154. doi: 10.1074/jbc.M110.178012

Wang, S., Ke, Y., Lu, G., Song, Z., Yu, L., Xiao, S., et al. (2013). Vasculogenic mimicry is a prognostic factor for postoperative survival in patients with glioblastoma. *J. Neurooncol.* 112, 339–345. doi: 10.1007/s11060-013-1077-7

Wang, Q., Xie, L., Dang, Y., Sun, X., Xie, T., Guo, J., et al. (2019). OSlms: A web server to evaluate the prognostic value of genes in Leiomyosarcoma. *Front. Oncol.* 9, 190. doi: 10.3389/fonc.2019.00190

Wang, Q., Wang, F., Lv, J., Xin, J., Xie, L., Zhu, W., et al. (2019). Interactive online consensus survival tool for esophageal squamous cell carcinoma prognosis analysis. *Oncol. Lett.* 18, 1199–1206. doi: 10.3892/ol.2019.10440

Wu, Z., Cai, L., Lin, S., Xiong, Z., Lu, J., Mao, Y., et al. (2013). High-mobility group box 2 is associated with prognosis of glioblastoma by promoting cell viability, invasion, and chemotherapeutic resistance. *Neuro Oncol.* 15, 1264–1275. doi: 10.1093/neuonc/not078

Wu, W., Hu, Q., Nie, E., Yu, T., Wu, Y., Zhi, T., et al. (2017). Hypoxia induces H19 expression through direct and indirect Hif-1α activity, promoting oncogenic effects in glioblastoma. *Sci. Rep.* 7, 45029. doi: 10.1038/srep45029

Xie, L., Wang, Q., Dang, Y., Ge, L., Sun, X., Li, N., et al. (2019a). OSkirc: a web tool for identifying prognostic biomarkers in kidney renal clear cell carcinoma. *Future Oncol* 15, 3103–3110. doi: 10.2217/fon-2019-0296

Xie, L., Dang, Y., Guo, J., Sun, X., Xie, T., Zhang, L., et al. (2019b). High KRT8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes (Basel)* 10. doi: 10.3390/genes10010036

Yan, J., Kong, L. Y., Hu, J., Gabrusiewicz, K., Dibra, D., Xia, X., et al. (2015). Autoantibodies against GLEA2 and PHF3 in glioblastoma: tumor-associated autoantibodies correlated with prolonged survival. *J. Natl. Cancer Inst.* 107 (8). doi: 10.1002/ijc.20929

Yang, P., Cai, J., Yan, W., Zhang, W., Wang, Y., Chen, B., et al. (2016). Classification based on mutations of TERT promoter and IDH characterizes subtypes in grade II/III gliomas. *Neuro Oncol.* 18, 1099–1108. doi: 10.1093/neuonc/now021

Zhang, W., Chen, H., Lv, S., and Yang, H. (2016a). High CD133 expression is associated with worse prognosis in patients with glioblastoma. *Mol. Neurobiol.* 53, 2354–2360. doi: 10.1007/s12035-015-9187-1

Zhang, J., Chen, Y., Lin, G., Zhang, J., Tang, W., Huang, J., et al. (2016b). High IFIT1 expression predicts improved clinical outcome, and IFIT1 along with MGMT more accurately predicts prognosis in newly diagnosed glioblastoma. *Hum. Pathol.* 52, 136–144. doi: 10.1016/j.humpath.2016.01.013

Zhang, W., Sun, Y., Liu, L., and Li, Z. (2017). Prognostic significance of TNFR-associated factor 1 and 2 (TRAF1 and TRAF2) in glioblastoma23,4506–4512. *Med. Sci. Monit.* 23, 4506–4512. doi: 10.12659/msm.903397

Zhang, G., Wang, Q., Yang, M., Yuan, Q., Dang, Y., Sun, X., et al. (2019). OSblca: A web server for investigating prognostic biomarkers of bladder cancer patients. *Front. Oncol.* 9, 466. doi: 10.3389/fonc.2019.00466

Zhao, J., Tang, H., Zhao, H., Che, W., Zhang, L., and Liang, P. (2015). SEMA6A is a prognostic biomarker in glioblastoma. *Tumour Biol.* 36, 8333–8340. doi: 10.1007/s13277-015-3584-y

Zupancic, K., Blejec, A., Herman, A., Veber, M., Verbovsek, U., Korsic, M., et al. (2014). Identification of plasma biomarker candidates in glioblastoma using an antibody-array-based proteomic approach. *Radiol. Oncol.* 48, 257–266. doi: 10.2478/raon-2014-0014

Conflict of Interest: Author MZ is employed by company of Nanjing Jiliang Biotechnology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fgene.2020.00253

Integrated Analysis to Evaluate the Prognostic Value of Signature mRNAs in Glioblastoma Multiforme

Ji'an Yang, Long Wang, Zhou Xu, Liquan Wu, Baohui Liu, Junmin Wang, Daofeng Tian, Xiaoxing Xiong and Qianxue Chen**

Department of Neurosurgery, Renmin Hospital of Wuhan University, Wuhan, China

OPEN ACCESS

Edited by:
Wan Zhu,
Stanford University, United States

Reviewed by:
Fuhai Li,
Washington University in St. Louis, United States
Alfred Grant Schissler,
University of Nevada, Reno, United States

****Correspondence:***
Xiaoxing Xiong
xiaoxingxiong@whu.edu.cn
Qianxue Chen
chenqx666@whu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *04 November 2019*
Accepted: *02 March 2020*
Published: *31 March 2020*

Citation:
Yang J, Wang L, Xu Z, Wu L, Liu B, Wang J, Tian D, Xiong X and Chen Q (2020) Integrated Analysis to Evaluate the Prognostic Value of Signature mRNAs in Glioblastoma Multiforme. Front. Genet. 11:253. doi: 10.3389/fgene.2020.00253

Background: Gliomas are the most common intracranial tumors and are classified as I–IV. Among them, glioblastoma multiforme (GBM) is the most common invasive glioma with a poor prognosis. New molecular biomarkers that can predict clinical outcomes in GBM patients must be identified, which will help comprehend their pathogenesis and supply personalized treatment. Our research revealed four powerful survival indicators in GBM by reanalyzing microarray data and genetic sequencing data in public databases. Moreover, it unraveled new potential therapeutic targets which could help improve the survival time and quality of life of GBM patients.

Materials and Methods: To identify prognostic signatures in GBMs, we analyzed the gene profiling data of GBM and standard brain samples from the Gene Expression Omnibus, including four datasets and RNA sequencing data from The Cancer Genome Atlas (TCGA) containing 152 glioblastoma tissues. We performed the differential analysis, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, weighted gene co-expression network analysis (WGCNA) and Cox regression analysis.

Results: After differential analysis in GSE12657, GSE15824, GSE42656 and GSE50161, overlapping differentially expressed genes were identified. We identified 110 up-regulated DEGs and 75 down-regulated DEGs in the GBM samples. Significantly enriched subclasses of the GO classification of these genes included mitotic sister chromatid separation, mitotic nuclear division and so on. In KEGG pathway analysis, the most abundant terms were ECM-receptor interaction and protein digestion and absorption. WGCNA analysis was performed on these 185 DEGs in 152 glioblastoma samples obtained from TCGA, and gene co-expression networks were constructed. We then performed a multivariate Cox analysis and established a Cox proportional hazards regression model using the top 20 genes significantly correlated with survival time. We

identified a four-protein prognostic signature that could divide patients into high-risk and low-risk groups. Increased expression of SLC12A5, CCL2, IGFBP2, and PDPN was associated with increased risk scores. Finally, the K-M curves confirmed that these genes could be used as independent predictors of survival in patients with glioblastoma.

Conclusion: Our analytical study identified a set of potential biomarkers that could predict survival and may contribute to successful treatment of GBM patients.

Keywords: glioblastoma, GEO, TCGA, WGCNA, prognosis biomarkers

INTRODUCTION

Gliomas are the most common intracranial tumors and are classified as grades I–IV according to World Health Organization (WHO) Classification of Tumors of the Central Nervous System (CNS). Among them, glioblastoma multiforme (GBM) is the most common primary brain tumor in adults with a poor prognosis (Reni et al., 2017). Patients with glioblastoma multiforme usually survive for less than 15 months after diagnosis and treatment. Therefore, it is crucial to develop appropriate and effective biomarkers to predict the prognosis of patients with glioblastoma. Various tumor related biomarkers have been found in glioblastoma, including epidermal growth factor receptor (EGFR), mutant form of the EGFR (EGFRvIII), vascular endothelial growth factor (VEGF), p53 and Phosphate and tensin homolog deleted on chromosome 10 (PTEN), Retinoblastoma (RB1) and Isocitrate dehydrogenase (IDH) (Appin and Brat, 2015). Some of these markers can predict therapeutic effect and clinical prognosis (Garrett-Bakelman and Melnick, 2013; Network, 2013; Westphal and Lamszus, 2015). Methylation status of the promoter of O-6-methylguanine-DNA methyltransferase (MGMT) is related to the sensitivity of temozolamide therapy and the prognosis of patients (Hegi et al., 2005; Wang et al., 2018). Loss of heterozygosity (LOH) of 1p/19q is another prognostic indicator, representing a better prognosis (Wiestler et al., 2014; Zhao et al., 2014). However, these markers can only be applied to specific parts of glioblastoma patients, and their proportion is not high. It is still necessary to identify novel molecular biomarkers that can predict the clinical outcome of GBM patients, which could help comprehend their pathogenesis and supply personalized treatment.

With the rapid development of sequencing technology and bioinformatics, they have provided new ideas for the study of clinical problems and related pathological mechanisms of various cancers. The Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA) and other public databases are broadly integrated collections of microarray data and gene sequencing data, enabling investigators to perform systematic analysis, which can help improve the diagnostic methods and survival prognosis of cancer patients. Considering different detection methods used by different technological platforms, as shown in **Figure 1**, various data processing and analysis methods are being explored. In this study, the RobustRankAggreg (RRA) (Kolde et al., 2012) method was used to combine the results of several separate studies to improve statistical power. Meanwhile, weighted gene co-expression network analysis (WGCNA) (Fuller et al., 2007; Langfelder and Horvath, 2008) was adopted to construct free-scale gene co-expression networks to identify core genes associated with clinical outcomes. These core genes may have important clinical significance and can be used as diagnostic and prognostic biomarkers or therapeutic targets.

MATERIALS AND METHODS

Microarray Data

Gene profiling data of GBM and normal brain samples were downloaded from the GEO[1], a public functional genomics data repository. Four datasets were selected for bioinformatics analysis, including GSE12657 (GPL8300, Affymetrix Human Genome U95 Version 2 Array),GSE50161 (GPL570, Affymetrix Human Genome U133 Plus 2.0 Array) (Griesinger et al., 2013), GSE42656 (GPL6947, Illumina HumanHT-12 V3.0 expression chip) (Henriquez et al., 2013) and GSE15824 (GPL570, Affymetrix Human Genome U133 Plus 2.0 Array) (Grzmil et al., 2011). All raw data were downloaded from the GEO database.

Microarray Data Normalization and Probe Annotation

The microarray data were quantile normalized using the "limma" package (Ritchie et al., 2015). After the data were normalized, the probe data in the original format were mapped to the gene symbols based on the annotation information. If multiple probes correspond to a gene, the average expression value of these probes was calculated as the expression of the gene (Xu et al., 2018). For probes with missing values, the "impute" package[2] was used to fill in missing values.

Download and Pre-processing of RNA-seq Data From TCGA

RNA sequencing data of human glioblastoma samples were available from the TCGA data portal[3], which contained 152 glioblastoma tissues. These data were then constructed into a matrix of RNA sequences, where gene symbols were rows and patient barcodes were column names. The clinical

[1]http://www.ncbi.nlm.nih.gov/geo

[2]http://bioconductor.org/packages/release/bioc/html/impute.html

[3]https://cancergenome.nih.gov/

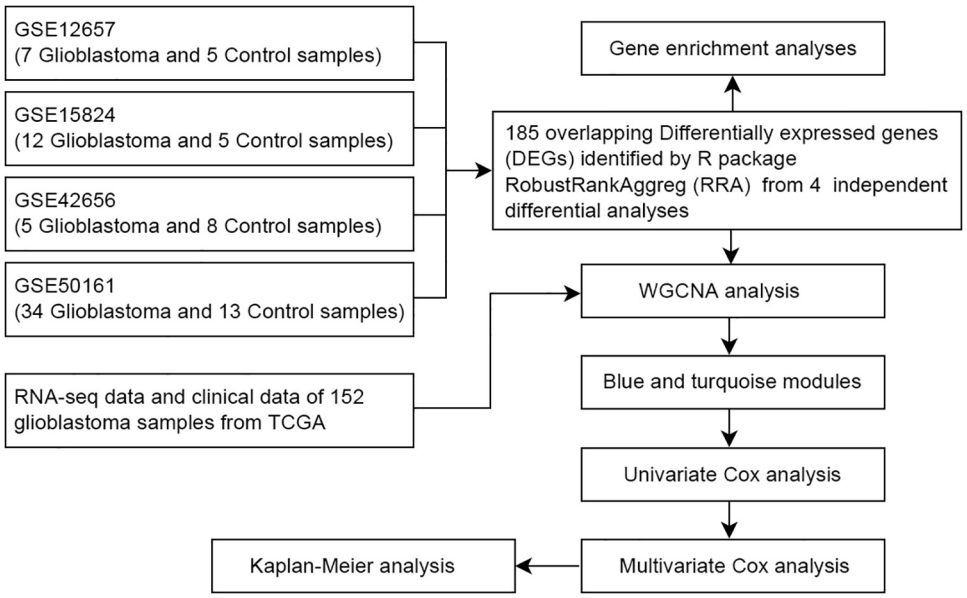


FIGURE 1 | Flow chart of data collection and analysis.

metadata of 152 samples were also downloaded and filtered for useful information.

Differential Analysis

Difference analysis was performed on four GEO datasets using the R package "limma" (Ritchie et al., 2015). In order to determine the best ranking results of the differential genes, a new robust rank aggregation method was used, which was implemented as the R package "RobustRankAggreg" (RRA)[4] (Kolde et al., 2012).

GO and KEGG Enrichment Analysis

The enrichment analysis of the KEGG pathway and Gene Ontology terms were performed through the R package "clusterProfiler"[5] (Yu et al., 2012; Yu et al., 2015). Enriched ontological terms and pathways ($P < 0.05$) were visualized as histograms.

Weighted Gene Co-expression Network Analysis

The R software package "WGCNA" was used for weighted gene co-expression network analysis (Langfelder and Horvath, 2008). It is an algorithm for constructing co-expression networks, defined by the similarity of gene co-expression. First, we calculated the Pearson correlation between each pair of differential genes and obtained a similarity matrix (sij). Second, the similarity matrix was converted into an adjacency matrix. The topological matrix was created using topological overlap measure (TOM) (Yip and Horvath, 2007). Finally, we chose the Dynamic hybrid cut method to identify co-expression gene modules (Langfelder et al., 2008). Details on the algorithm were available on request.

Cox Regression Analysis

To validate the significance of the prognostic risk genes screened above, we used univariate Cox proportional hazards regression to assess the effect of expression of these genes on survival time in GBM patients. Limited to the strength of computer calculation, we used the top 20 genes significantly related to survival time to perform the multivariate Cox analysis. Then, statistically significant genes were used to construct a multivariate cox regression model. The above analysis had used the R package "survival"[6] (Therneau and Grambsch, 2000). The R package "survivalROC"[7] was used to perform the receiver operating characteristic curve (ROC) to evaluate the accuracy of the model (Heagerty et al., 2000).

Statistical Analysis

All statistical tests and charts were performed using RStudio. $P < 0.05$ was considered statistically significant. These graphics were then integrated and displayed using Photoshop.

RESULTS

Screening for Differentially Expressed Genes (DEGs)

The differential analysis in GSE12657, GSE50161, GSE42656, and GSE15824 was performed by "limma" algorithm. Subsequently, 185 overlapping differentially expressed genes were identified by

[4]https://CRAN.R-project.org/package=RobustRankAggreg
[5]https://github.com/YuLab-SMU/clusterProfiler
[6]https://CRAN.R-project.org/package=survival
[7]https://CRAN.R-project.org/package=survivalROC

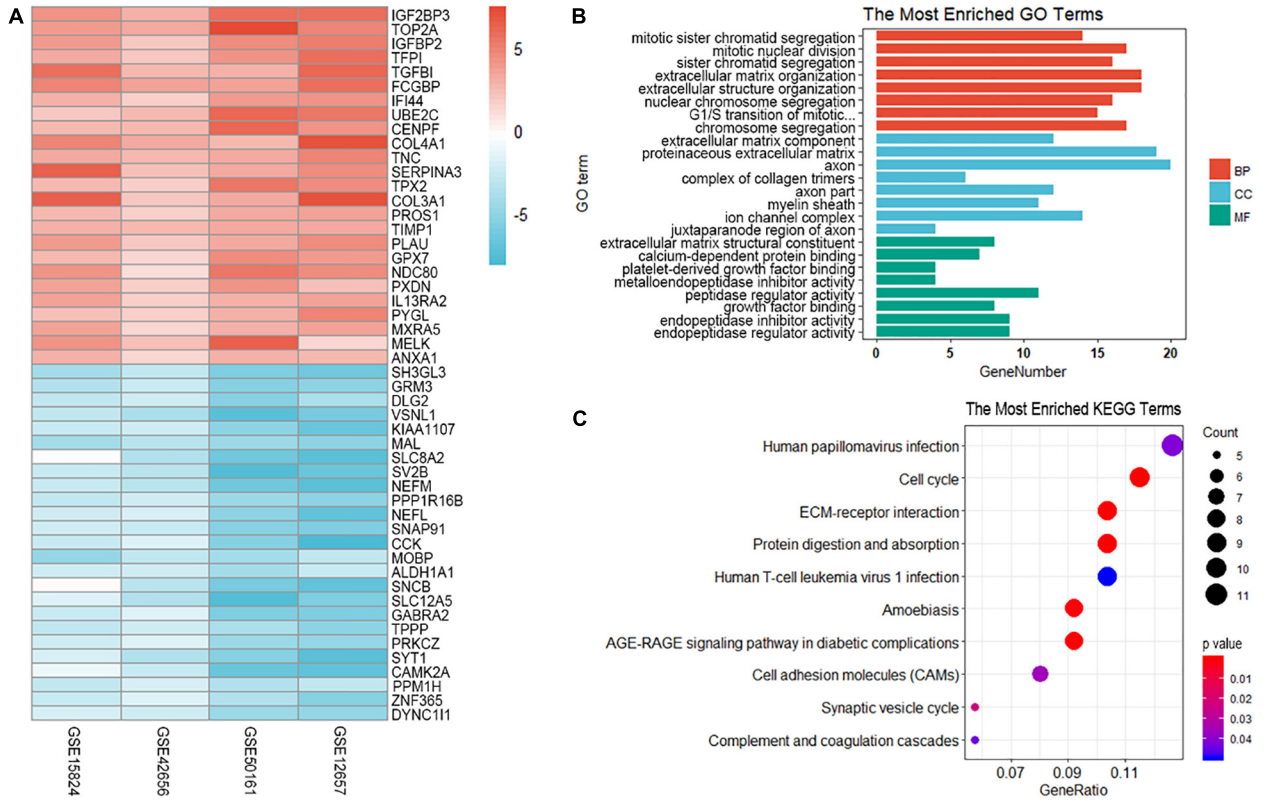


FIGURE 2 | Visualization and enrichment analysis of differentially expressed genes. **(A)** The heatmap of the top 50 DEGs. **(B)** GO enrichment analysis of differentially expressed genes. BP, biological process; CC, cellular component; MF, molecular function; **(C)** KEGG enrichment analysis of differentially expressed genes. The size of the dot represents the count number of genes in one KEGG term.

"RobustRankAggreg," of which 110 were up-regulated and 75 were down-regulated in GBM samples. The top 50 DEGs were visualized as heatmap (**Figure 2A**).

GO and KEGG Enrichment Analysis of DEGs

To explore the biological relevance of DEGs, Gene Ontology (Ashburner et al., 2000) and KEGG (Ogata et al., 2000) pathway enrichment analyses were performed. GO and KEGG analysis predicted that these genes were involved in several important physiological processes. These genes were significantly enriched in the following subclasses of GO classification:mitotic sister chromatid segregation (GO: 0000070 P = 2.67E-10), mitotic nuclear division (GO: 0140014 1.28E – 09), sister chromatid segregation (GO: 0000819 P = 2.44E – 09), extracellular matrix component (GO: 0044420 P = 2.69E – 09),proteinaceous extracellular matrix (GO: 0005578 P = 4.39E – 09) and extracellular matrix structural constituent (GO: 0005201 P = 1.57E – 06). The KEGG pathway analysis showed that the most enriched terms were ECM-receptor interaction (hsa04512 P = 4.18E – 07), protein digestion and absorption (hsa04974, P = 9.33E – 07) (**Figures 2B,C**).

Co-expression Network Construction and Visualization

Afterward, the WGCNA analysis was performed to construct gene co-expression networks. We analyzed the 185 DEGs identified above in the data of 152 glioblastoma samples from TCGA and divided the 185 genes into three modules (**Figure 3**). The blue and turquoise co-expressed modules were identified to further analysis (**Figure 4A**). In order to explore whether different modules have different biological functions, enrichment analysis was also performed on the modules. It was found that the biological processes of the blue module mainly focused on cell proliferation and division. However, the turquoise module focused on signal molecule delivery (**Figure 4B**). Whereafter, the co-expression networks of the modules were exported into Cytoscape and visualized (Shannon et al., 2003). The nodes were defined as individual genes in the networks, and the edges were defined as the interactions between genes (**Figure 4C**).

Construction of the Cox Proportional Hazards Regression Model Based on Hub Genes and Kaplan–Meier Analysis

The selected DEGs were further used to perform univariate Cox analysis. We then performed a multivariate Cox analysis using the top 20 genes significantly correlated with

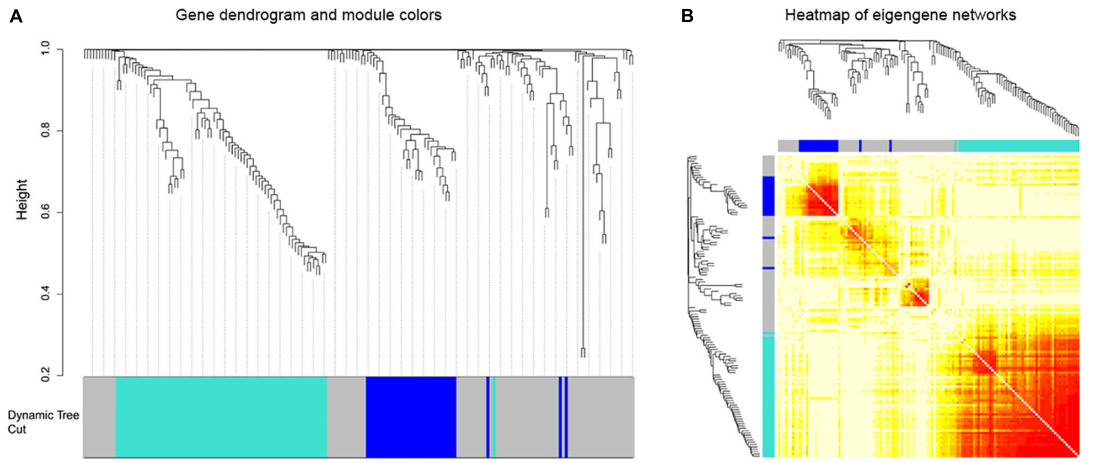


FIGURE 3 | Weighted gene co-expression network of glioblastoma. **(A)** Gene dendrograms obtained by average linkage hierarchical clustering of 185 genes based on consensus Topological Overlap with the corresponding module colors indicated by the color row. **(B)** The eigengene networks were shown as heatmap. The deeper the color expressed a high adjacency.

survival time, and constructed a Cox proportional hazards regression model from 152 patients with glioblastoma. Based on the above model, the following formula was used to calculate the risk score for predicting survival time: risk score = (0.2239*expression level of CCL2) + (0.3375*expression level of IGFBP2) + (0.1516*expression level of PDPN) + (0.2276*expression level of SLC12A5) (**Figure 5**). According to the median risk score, 152 patients were divided into high-risk (N = 76) and low-risk (N = 76) groups. The 5-year survival rate in the high-risk group was significantly lower than low-risk group. Increased expression of SLC12A5, CCL2, IGFBP2, and PDPN was associated with increased risk scores (**Figure 6A**). The area under the ROC curve was 0.701 (**Figure 6B**), indicating the high predictive value. Meanwhile, K-M curves confirmed that these three genes (CCL2, IGFBP2, and PDPN) could be used as independent predictors of survival in patients with glioblastoma (**Figures 6C–F**).

DISCUSSION

High-throughput microarray technology provides insights into pathogenesis, molecular heterogeneity and treatment response. The biological conclusions are inconsistent due to differences in detection platforms and laboratory protocols and noisy microarray data. To overcome these limitations, it is considerable to analyze these data set separately and then summarize different lists of results. In our research, we identified 185 DEGs for GBM derived from independent profiling datasets by applying "limma" algorithm and "RRA" method. This method using a probabilistic model probabilistic model makes the algorithm parameter free and robust to outliers, noise and errors, and facilitates the calculation of significance probabilities for all the elements in the final ranking. This strategy has been widely applied to identify disease-related genes (Kolde et al., 2012; Xiao, 2020; Xiong et al., 2018).

Subsequently, the WGCNA analysis was performed on RNA-seq data obtained from TCGA on those 185 DEGs to identify two co-expressed modules (blue and turquoise). WGCNA is a recently developed method to construct a weighted gene co-expression network and a new analytic approach to move beyond single-gene comparisons (Giulietti et al., 2018). The WGCNA algorithm has been used to identify disease-related genes, biological pathways and therapeutic targets for diseases such as familial combined hyperlipidemia, Osteoporosis, Autistic, and Alzheimer disease (Goh et al., 2007; He et al., 2011; Tang et al., 2017). It also has been used in neuroscience and oncology. Michael C Oldham performed the WGCNA in normal human brains to identify co-expressed gene modules that reflected the underlying cellular composition of brain tissue and system-level molecules related to neuroanatomy (Oldham et al., 2006). The large number of tumor RNA-seq data and other high-throughput data resources such as TCGA provide a broad opportunity for the application of WGCNA in cancer research. To date, there have been similar studies on gliomas. Zhou and colleagues revisited the gene expression profile data downloaded from GEO to identify novel genes associated with pediatric pilocytic astrocytoma using the WGCNA analysis. They identified nine network modules associated with pilocytic astrocytomas. The further functional analysis revealed that these genes were involved in the regulation of cell differentiation (Zhou and Man, 2016). S. Horvath used WGCNA to identify several gene co-expression modules and revealed abnormal spindle-like microcephaly-associated protein (ASPM) that might function as a potential molecular target in glioblastoma (Horvath et al., 2006). In addition, Upton A and his colleagues used the WGCNA algorithm and further identified 92 genes that were associated with different evolutionary stages of glioblastoma (Upton and Arvanitis, 2014). In our research, the biological processes of the blue module mainly focused on cell proliferation and division. While, the turquoise module focused on signal molecule delivery.

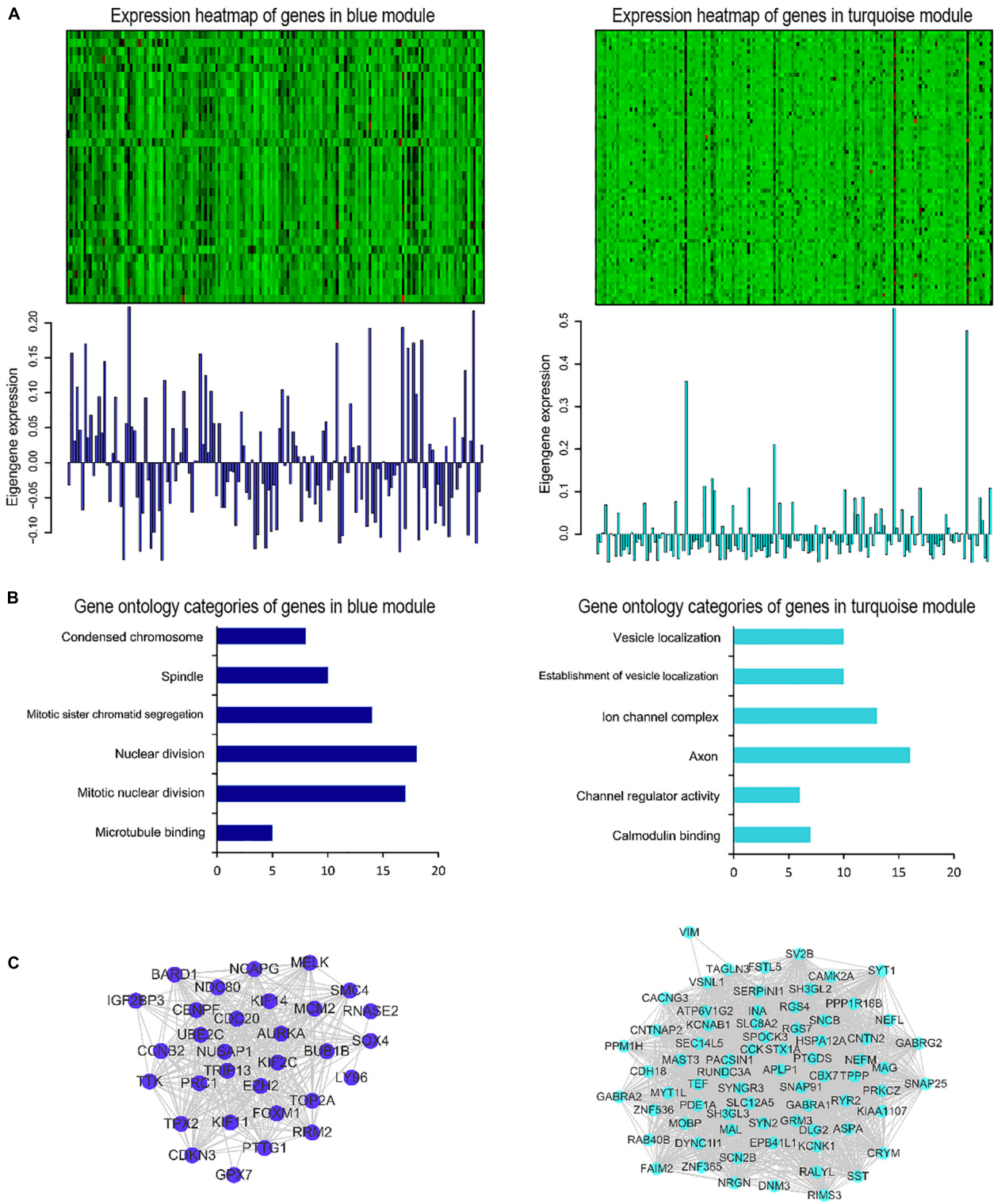


FIGURE 4 | Gene co-expression modules associated with glioblastoma. **(A)** Heatmap of genes belonging to the co-expression module. Corresponding module eigengene values (y-axis) across samples (x-axis). **(B)** Relevant gene ontology categories of enriched genes in the blue and turquoise modules. **(C)** Visualization of the gene co-expression network of the blue and turquoise modules.

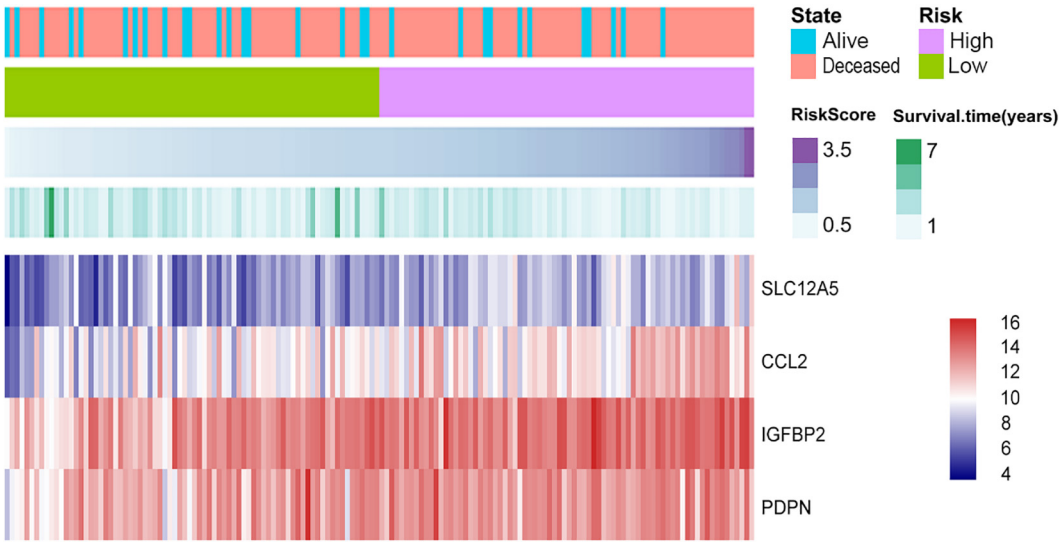


FIGURE 5 | Cox proportional hazards regression model. Purple depths of the third column reveal the risk score of the low-risk and high-risk groups. Green depths of the fourth column display the survival status and time of 152 glioblastomas. The lowest column shows the heatmap of the model genes.

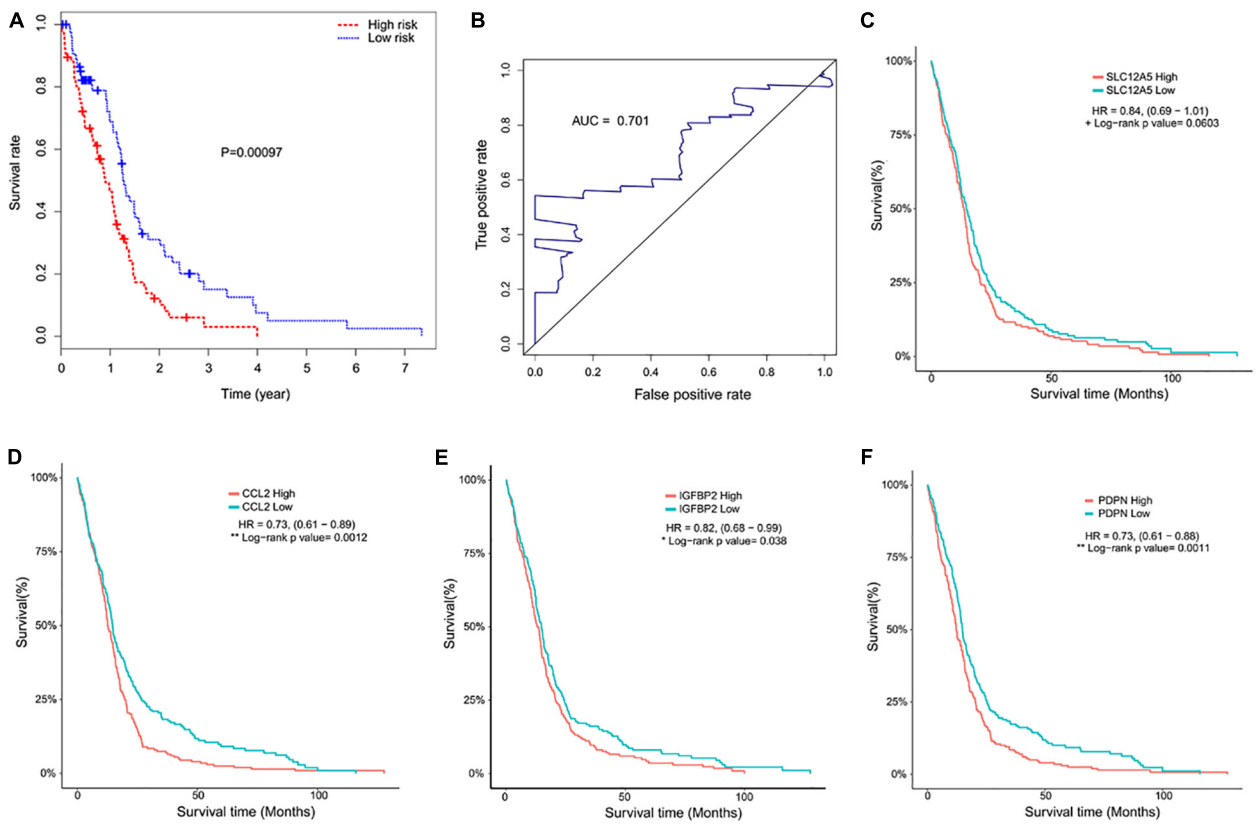


FIGURE 6 | Kaplan–Meier curves and receiver operating characteristic (ROC). **(A)** Kaplan–Meier curve showed that the mortality in the high-risk group was higher than that in the low risk group ($P < 0.001$). **(B)** Time-dependent ROC curve indicated a higher predictive value. The area under the ROC curve (AUC) was 0.701. **(C–F)** Kaplan–Meier curves of the four predictive indicators.

These results help to understand the occurrence and development of glioblastoma to some extent, and further research is needed.

Cox proportional hazards regression has been widely used to examine the prognostic value of candidate predictors in human diseases (Degnim et al., 2018; Liu et al., 2018). Aoki K used the Cox proportional hazards regression model to study the effects of genetic variation and clinicopathological factors on the survival of diffuse low-grade gliomas (LGGs). The authors reported subtype-specific genetic alterations could stratify patients with different LGG subtypes (Aoki et al., 2018). By constructing the Cox proportional hazards regression model, we selected an optimal four-gene model (SLC12A5 + CCL2 + IGFBP2 + PDPN) for prognosis prediction. Among the genes in this model, solute carrier family 12, member 5 (SLC12A5) was considered as a neuron marker, but it has not been reported in glioma-related studies. Chemokine ligand 2 (CCL2) is one of several cytokine genes and could be secreted by astrocytoma cells and myeloid cells. Importantly, CCL2 then recruits regulatory T cells (Tregs) and myeloid-derived suppressor cells (MDSCs) through CCR4 and CCR2 as significant contributors to the potently immunosuppressive glioma microenvironment (Carrillo-de et al., 2012; Braganhol et al., 2015; Chang et al., 2016; Lu et al., 2017). Overexpression of Insulin-like growth factor binding protein 2 (IGFBP2) has been reported to be involved in the progression of many types of cancer. In gliomas, IGFBP2 is considered to be an oncogene that causes glioma progression through integrin/ILK/NF-kB pathway (Phillips et al., 2016). According to reports, Podoplanin (PDPN) was a novel candidate gene that might play an essential role in glioblastoma pathogenesis and response to treatment (Sailer et al., 2013; Krishnan et al., 2018). However, these genes and the related signaling pathways and mechanisms involved are still not clear enough.

Our research has some limitations. First, in order to reduce intensity of computer operation, we used the top 20 genes significantly related to survival time to perform the multivariate Cox analysis. But constructing a model with more genes might get more meaningful results. Second, due to the lack of survival data in the GEO datasets, we did not validate the prognostic value of the four-gene model. Third, the expression levels of corresponding proteins have not been verified in tissue samples. Finally, we used the "RRA" method to identify DEGs, and in this process, the tumor heterogeneity might be ignored. We might lose some key genes and pathways in the development of gliomas in the integration analysis. In summary, in this study, we tried to apply a new procedure to screen out some new biomarkers that can help the diagnosis and treatment of glioblastoma. Although the methods are not new, combining them with new process may bring new perspectives. We identified a four-gene (SLC12A5 + CCL2 + IGFBP2 + PDPN) Cox proportional hazards regression model for prognosis prediction. Although the specific mechanism remains to be studied, these genes could be considered as risk factors for GBM patients and novel therapeutic targets.

DATA AVAILABILITY STATEMENT

Microarray data were retrieved from the GEO data repository (http://www.ncbi.nlm.nih.gov/geo/) with the accession numbers: GSE12657, GSE50161, GSE42656, and GSE15824. The RNA sequencing data of human glioblastoma samples were obtained from the TCGA data portal (https://portal.gdc.cancer.gov).

AUTHOR CONTRIBUTIONS

JY contributed to the publication search, data extraction, draft writing, and conception and design. LW, ZX, LqW, JW, DT, XX, and QC contributed to the quality assessment, conception and design, and editing. BL contributed to the statistical analysis.

FUNDING

The present study was supported by the National Natural Science Foundation of China (No. 81572489).

ACKNOWLEDGMENTS

The results in this research were based upon data from the Gene Expression Omnibus and The Cancer Genome Atlas established by the NCI and NHGRI. Information about GEO and TCGA and the investigators and institutions that constitute the GEO and TCGA research network can be found at http://www.ncbi.nlm.nih.gov/geo and http://cancergenome.nih.gov/.

REFERENCES

Aoki, K., Nakamura, H., Suzuki, H., Matsuo, K., Kataoka, K., Shimamura, T., et al. (2018). Prognostic relevance of genetic alterations in diffuse lower-grade gliomas. *Neuro Oncol.* 20, 66–77. doi: 10.1093/neuonc/nox132

Appin, C. L., and Brat, D. J. (2015). Biomarker-driven diagnosis of diffuse gliomas. *Mol. Aspects Med.* 45, 87–96. doi: 10.1016/j.mam.2015.05.002

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29.

Braganhol, E., Kukulski, F., Levesque, S. A., Fausther, M., Lavoie, E. G., Zanotto-Filho, A., et al. (2015). Nucleotide receptors control IL-8/CXCL8 and MCP-1/CCL2 secretions as well as proliferation in human glioma cells. *Biochim. Biophys. Acta* 1852, 120–130. doi: 10.1016/j.bbadis.2014.10.014

Carrillo-de, S. M., Gomez, A., Ros, C. M., Ros-Bernal, F., Martin, E. D., Perez-Valles, A., et al. (2012). CCL2-expressing astrocytes mediate the extravasation of T lymphocytes in the brain. Evidence from patients with glioma and experimental models in vivo. *PLoS One* 7:e30762. doi: 10.1371/journal.pone.0030762

Chang, A. L., Miska, J., Wainwright, D. A., Dey, M., Rivetta, C. V., Yu, D., et al. (2016). CCL2 Produced by the glioma microenvironment is essential for the recruitment of regulatory T Cells and myeloid-derived suppressor cells. *Cancer Res.* 76, 5671–5682. doi: 10.1158/0008-5472.can-16-0144

Degnim, A. C., Winham, S. J., Frank, R. D., Pankratz, V. S., Dupont, W. D., Vierkant, R. A., et al. (2018). Model for predicting breast cancer risk in women with atypical hyperplasia. *J. Clin. Oncol.* 36, 1840–1846. doi: 10.1200/JCO.2017.75.9480

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome* 18, 463–472. doi: 10.1007/s00335-007-9043-3

Garrett-Bakelman, F. E., and Melnick, A. M. (2013). Differentiation therapy for IDH1/2 mutant malignancies. *Cell Res.* 23, 975–977. doi: 10.1038/cr.2013.73

Giulietti, M., Occhipinti, G., Righetti, A., Bracci, M., Conti, A., Ruzzo, A., et al. (2018). Emerging biomarkers in bladder cancer identified by network analysis of transcriptomic data. *Front. Oncol.* 8:450. doi: 10.3389/fonc.2018.00450

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690.

Griesinger, A. M., Birks, D. K., Donson, A. M., Amani, V., Hoffman, L. M., Waziri, A., et al. (2013). Characterization of distinct immunophenotypes across pediatric brain tumor types. *J. Immunol.* 191, 4880–4888. doi: 10.4049/jimmunol.1301966

Grzmil, M., Morin, P. J., Lino, M. M., Merlo, A., Frank, S., Wang, Y., et al. (2011). MAP kinase-interacting kinase 1 regulates SMAD2-dependent TGF-beta signaling pathway in human glioblastoma. *Cancer Res.* 71, 2392–2402. doi: 10.1158/0008-5472.CAN-10-3112

He, D., Liu, Z. P., and Chen, L. (2011). Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics* 12:592. doi: 10.1186/1471-2164-12-592

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56, 337–344. doi: 10.1111/j.0006-341x.2000.00337.x

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., de Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003.

Henriquez, N. V., Forshew, T., Tatevossian, R., Ellis, M., Richard-Loendt, A., Rogers, H., et al. (2013). Comparative expression analysis reveals lineage relationships between human and murine gliomas and a dominance of glial signatures during tumor propagation in vitro. *Cancer Res.* 73, 5834–5844. doi: 10.1158/0008-5472.CAN-13-1299

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17402–17407. doi: 10.1073/pnas.0608396103

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 573–580. doi: 10.1093/bioinformatics/btr709

Krishnan, H., Rayes, J., Miyashita, T., Ishii, G., Retzbach, E. P., Sheehan, S. A., et al. (2018). Podoplanin: an emerging cancer biomarker and therapeutic target. *Cancer Sci.* 109, 1292–1299. doi: 10.1111/cas.13580

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416. doi: 10.1016/j.cell.2018.02.052

Lu, B., Zhou, Y., Su, Z., Yan, A., and Ding, P. (2017). Effect of CCL2 siRNA on proliferation and apoptosis in the U251 human glioma cell line. *Mol. Med. Rep.* 16, 3387–3394. doi: 10.3892/mmr.2017.6995

Network, T. C. (2013). Corrigendum: comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 494:506. doi: 10.1038/nature11903

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34.

Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17973–17978. doi: 10.1073/pnas.060593 8103

Phillips, L. M., Zhou, X., Cogdell, D. E., Chua, C. Y., Huisinga, A., Hess, R. K., et al. (2016). Glioma progression is mediated by an addiction to aberrant IGFBP2 expression and can be blocked using anti-IGFBP2 strategies. *J. Pathol.* 239, 355–364. doi: 10.1002/path.4734

Reni, M., Mazza, E., Zanon, S., Gatta, G., and Vecht, C. J. (2017). Central nervous system gliomas. *Crit. Rev. Oncol. Hematol.* 113, 213–234. doi: 10.1016/j.critrevonc.2017.03.021

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Sailer, M. H., Gerber, A., Tostado, C., Hutter, G., Cordier, D., Mariani, L., et al. (2013). Non-invasive neural stem cells become invasive in vitro by combined FGF2 and BMP4 signaling. *J. Cell Sci.* 126(Pt 16), 3533–3540. doi: 10.1242/jcs.125757

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Tang, R. X., Chen, W. J., He, R. Q., Zeng, J. H., Liang, L., Li, S. K., et al. (2017). Identification of a RNA-Seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. *Oncotarget* 8, 50761–50773. doi: 10.18632/oncotarget.17098

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer.

Upton, A., and Arvanitis, T. N. (2014). Using evolutional properties of gene networks in understanding survival prognosis of glioblastoma. *IEEE J. Biomed. Health Inform.* 18, 810–816. doi: 10.1109/JBHI.2013.2282569

Wang, W., Zhao, Z., Wu, F., Wang, H., Wang, J., Lan, Q., et al. (2018). Bioinformatic analysis of gene expression and methylation regulation in glioblastoma. *J. Neurooncol.* 136, 495–503. doi: 10.1007/s11060-017-2688-1

Westphal, M., and Lamszus, K. (2015). Circulating biomarkers for gliomas. *Nat. Rev. Neurol.* 11, 556–566. doi: 10.1038/nrneurol.2015.171

Wiestler, B., Capper, D., Sill, M., Jones, D. T., Hovestadt, V., Sturm, D., et al. (2014). Integrated DNA methylation and copy-number profiling identify three clinically and biologically relevant groups of anaplastic glioma. *Acta Neuropathol.* 128, 561–571. doi: 10.1007/s00401-014-1315-x

Xiao, Y. (2020). Construction of a circRNA-miRNA-mRNA network to explore the pathogenesis and treatment of pancreatic ductal adenocarcinoma. *J. Cell. Biochem.* 121, 394–406. doi: 10.1002/jcb.29194

Xiong, D. D., Dang, Y. W., Lin, P., Wen, D. Y., He, R. Q., Luo, D. Z., et al. (2018). A circRNA-miRNA-mRNA network identification for exploring underlying pathogenesis and therapy strategy of hepatocellular carcinoma. *J. Transl. Med.* 16:220. doi: 10.1186/s12967-018-1593-5

Xu, P., Yang, J., Liu, J., Yang, X., Liao, J., Yuan, F., et al. (2018). Identification of glioblastoma gene prognosis modules based on weighted gene co-expression network analysis. *BMC Med. Genomics* 11:96. doi: 10.1186/s12920-018-0407-1

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8:22. doi: 10.1186/1471-2105-8-22

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, G., Wang, L. G., Yan, G. R., and He, Q. Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Zhao, J., Ma, W., and Zhao, H. (2014). Loss of heterozygosity 1p/19q and survival in glioma: a meta-analysis. *Neuro Oncol.* 16, 103–112. doi: 10.1093/neuonc/not145

Zhou, R., and Man, Y. (2016). Integrated analysis of DNA methylation profiles and gene expression profiles to identify genes associated with pilocytic astrocytomas. *Mol. Med. Rep.* 13, 3491–3497. doi: 10.3892/mmr.2016.4943

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 April 2020
doi: 10.3389/fgene.2020.00302

Analysis of the Interaction Network of Hub miRNAs-Hub Genes, Being Involved in Idiopathic Pulmonary Fibers and Its Emerging Role in Non-small Cell Lung Cancer

Dong Hu Yu[1†], Xiao-Lan Ruan[2†], Jing-Yu Huang[3], Xiao-Ping Liu[1], Hao-Li Ma[1,4], Chen Chen[1,4], Wei-Dong Hu[3] and Sheng Li[1,4]*

[1] Department of Biological Repositories, Zhongnan Hospital, Wuhan University, Wuhan, China, [2] Department of Hematology, Renmin Hospital, Wuhan University, Wuhan, China, [3] Department of Thoracic Surgery, Zhongnan Hospital, Wuhan University, Wuhan, China, [4] Human Genetics Resource Preservation Center, Wuhan University, Wuhan, China

Idiopathic pulmonary fibrosis (IPF) is a fibrotic interstitial lung disease with lesions confined to the lungs. To identify meaningful microRNA (miRNA) and gene modules related to the IPF progression, GSE32537 (RNA-sequencing data) and GSE32538 (miRNA-sequencing data) were downloaded and processed, and then weighted gene co-expression network analysis (WGCNA) was applied to construct gene co-expression networks and miRNA co-expression networks. GSE10667, GSE70866, and GSE27430 were used to make a reasonable validation for the results and evaluate the clinical significance of the genes and the miRNAs. Six hub genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) and seven hub miRNAs (hsa-let-7b-5p, hsa-miR-26a-5p, hsa-miR-25-3p, hsa-miR-29c-3p, hsa-let-7c-5p, hsa-miR-29b-3p, and hsa-miR-26b-5p) were clarified and validated. Meanwhile, iteration network of hub miRNAs-hub genes was constructed, and the emerging role of the network being involved in non-small cell lung cancer (NSCLC) was also analyzed by several webtools. The expression levels of hub genes were different between normal lung tissues and NSCLC tissues. Six genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) and three miRNAs (hsa-miR-29c-3p, hsa-let-7c-5p, and hsa-miR-29b-3p) were related to the survival time of lung adenocarcinoma (LUAD). The interaction network of hub miRNAs-hub genes might provide common mechanisms involving in IPF and NSCLC. More importantly, useful clues were provided for clinical treatment of both diseases based on novel molecular advances.

Keywords: idiopathic pulmonary fibers, non-small cell lung cancer, weighted gene co-expression network analysis, hub genes, hub miRNAs, interaction network

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Jun Zhong,
National Cancer Institute (NCI), United States
Xiaoxi Zeng,
Sichuan University, China

***Correspondence:**
Sheng Li
lisheng-znyy@whu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *08 June 2019*
Accepted: *13 March 2020*
Published: *02 April 2020*

Citation:
Yu DH, Ruan X-L, Huang J-Y, Liu X-P, Ma H-L, Chen C, Hu W-D and Li S (2020) Analysis of the Interaction Network of Hub miRNAs-Hub Genes, Being Involved in Idiopathic Pulmonary Fibers and Its Emerging Role in Non-small Cell Lung Cancer. Front. Genet. 11:302. doi: 10.3389/fgene.2020.00302

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a chronic phlogistic interstitial lung disease with excessive tissue scarring and loss of function, and most patients with IPF would die of organ failure eventually (Datta et al., 2011; Lehtonen et al., 2016). To assess disease progression for the patients with IPF, the scores of St. George's Respiratory Questionnaire (SGRQ) are usually used, which have

a strong correlation with lung function significantly (Swigris et al., 2014, 2018; Lawrence et al., 2017). Besides, non-small cell lung cancer (NSCLC), which can mainly be categorized into lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), is commonly altering the course and mortality of IPF (Ballester et al., 2019). IPF and NSCLC are coexistent and affect each other, and majority of studies have shown that LUSC is the most frequent type of NSCLC in IPF patients, while LUAD is the second most frequent (Lee et al., 2014; Tomassetti et al., 2015; Kato et al., 2018). Studies have shown that the risk of NSCLC is higher in IPF patients, and it was reported that the cumulative prevalence of NSCLC is increased from IPF diagnosis (Kinoshita and Goto, 2019). Recent Studies indicated that the occurrence of IPF and NSCLC share the same genetic mutations and abnormal activation of signal pathways, suggesting potential molecular mechanisms between IPF and NSCLC, and there is speculation IPF could lead to cancer (Han et al., 2019; Kinoshita and Goto, 2019). IPF, which has a poor prognosis and a course that is unpredictable, thus needs for a more complete understanding of its mechanisms, and further research for IPF-NSCLC pathogenesis is also urgently needed.

MicroRNA (miRNA) is a class of gene regulator, and it can repress the expression of target genes by binding to the mRNAs (Taganov et al., 2007). In recent years, increasing evidences have revealed that multiple miRNAs can play as potential biomarkers for the prediction of IPF, including miR-92a (Berschneider et al., 2014), miR-let-7d (Huleihel et al., 2014), and miR-98 (Gao et al., 2014). However, studies of single miRNA cannot meet the requirement for exploration of IPF progression. miRNAs–mRNAs constitute networks, which are involved in many important cellular pathways, are badly needed to clarify exact mechanisms.

Though Fan has reported differently expressed genes and differently expressed miRNAs between normal tissue and IPF tissues (Fan et al., 2017), the relationships between hub genes and important clinical traits, hub miRNAs, and important clinical traits had not been rigorously studied. The weighted gene co-expression network analysis (WGCNA), which provides an effective way to explore the mechanisms behind certain traits, can solve this problem elegantly (Langfelder and Horvath, 2008). To fulfill these gaps, gene co-expression networks and miRNA co-expression networks were constructed by WGCNA to identify the gene and miRNA modules related to the scores of SGRQ in IPF, and the relationships between genes and miRNAs were predicted to construct miRNA–gene network, which would provide more information about the mechanisms of IPF progression, even IPF-NSCLC pathogenesis.

MATERIALS AND METHODS

Data Collection and Processing

A brief workflow for this study is indicated in **Figure 1**. Selection criteria on the Gene Expression Omnibus (GEO) database[1] are: (1) The datasets contain miRNA expression profiles and gene expression profiles; (2) there are normal group (normal tissue samples) and IPF group (IPF tissue samples) in the datasets; and (3) the number of samples in each group is more than 10. miRNA expression profiles (GSE32538 and GSE27430) and gene expression profiles (GSE32537, GSE10667, and GSE70866) related to IPF were downloaded from GEO database. All datasets were normalized with quantile normalization. The data quality was evaluated, and boxplot was used to compare before and after being standardized. The details of these datasets are listed in **Supplementary Table S1**. Among them, GSE32537 and GSE32538 were used to identify hub genes and hub miRNAs by WGCNA separately. After doing analysis of variance for GSE32537, we chose the top 25% most variant genes (2987 genes) for constructing networks, while we did not to do pretreatment for GSE32538 due to the small number of miRNAs (1801 miRNAs).

Construction of Co-expression Networks

Weighted gene co-expression network analysis was used to construct gene co-expression networks and miRNA co-expression networks (Langfelder and Horvath, 2008). The processes for constructing gene co-expression networks and miRNA co-expression networks were similar. So, we took the construction of weighted gene co-expression networks as an example. First, a matrix of similarity was constructed by calculating the correlations of the processed genes. Second, an appropriate power of β was chosen as the soft-thresholding parameter to construct a scale-free network. Third, the adjacency was transformed into a topological overlap matrix (TOM) by using TOM similarity, and the corresponding dissimilarity (1-TOM) was figured and the dissimilarity of module eigengenes (MEs) was estimated. Fourth, the genes with similar expression levels were categorized into the same module by DynamicTreeCut algorithm.

Identification of Clinically Significant Modules

The clinical trait that we concerned was the scores of SGRQ in IPF patients and key modules needed to be found in two networks separately. Above all, we worked out the relationship between clinical phenotype and MEs. MEs were deemed to represent the expression levels of all genes or miRNAs in the related module. In addition, mediated *p*-value of each gene or miRNA was calculated and then we worked out gene significance or miRNA significance (GS = lg P). Finally, we selected the most clinically significant module according to module significance (MS), which was the average GS of genes or miRNAs involved in the related module.

Functional and Pathway Enrichment Analysis

The Database for Annotation, Visualization and Integrate Discovery5 (DAVID)[2] is a database for several kinds of functional annotation (Huang et al., 2009). With the help of DAVID, we identified biological meaning of the genes in a given module

[1] https://www.ncbi.nlm.nih.gov/geo/

[2] https://david.ncifcrf.gov/

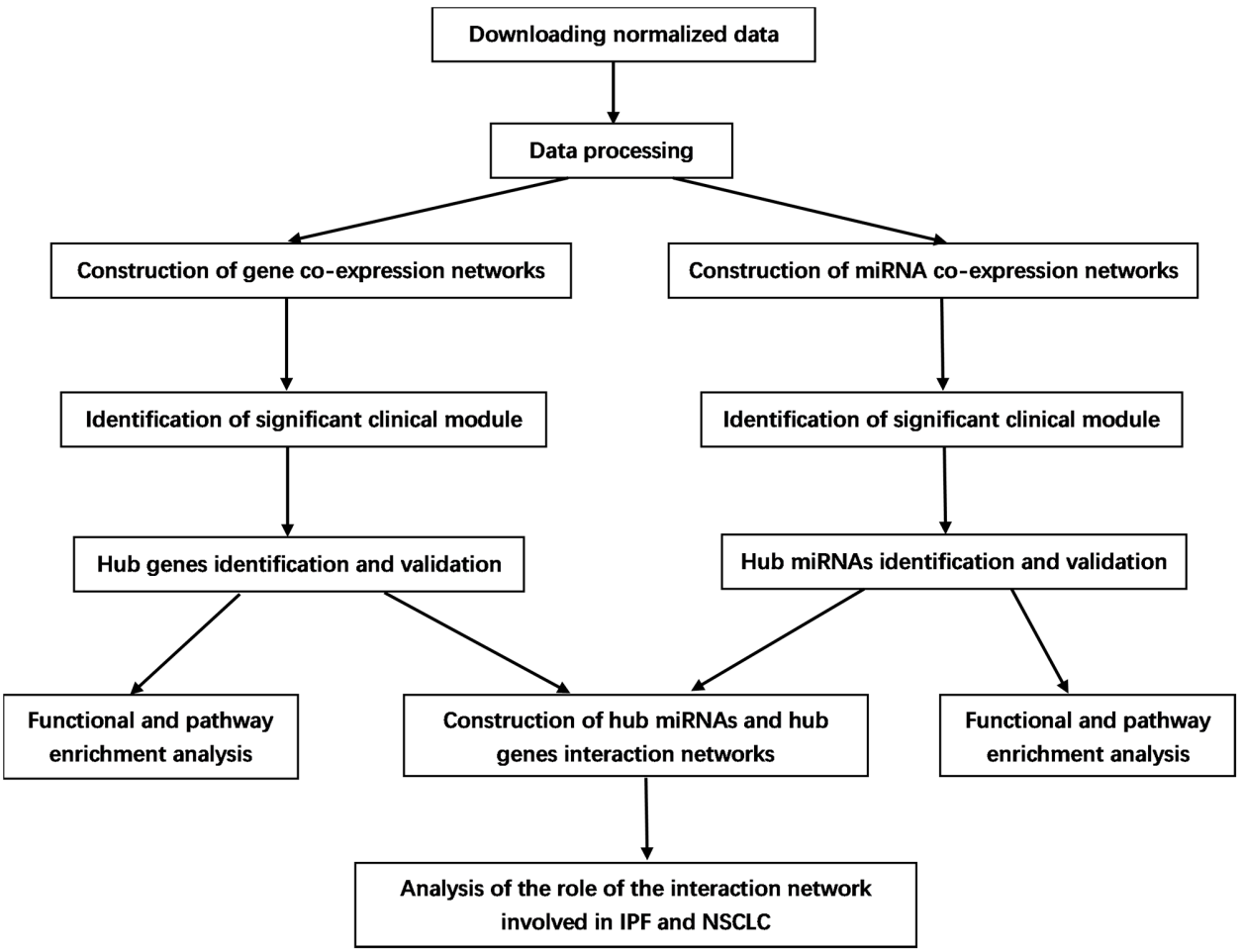


FIGURE 1 | Flow chart of data preparation, processing, analysis, and validation.

according to false discovery rate (FDR) < 0.05. GO includes three terms: biological process (BP), cellular component (CC), and molecular function (MF). Besides, GO (BP, CC, MF) and KEGG enrichment analyses for the miRNAs in the selected module were conducted using mirPath v.3, an online tool for miRNA pathway analysis (Vlachos et al., 2015).

Identification and Validation of Hub Genes and Hub miRNAs in IPF

The connectivity of module can be measured by absolute value of the Pearson's correlation. Besides, the relationship between clinical trait and genes can be measured by absolute value of the Pearson's correlation. The genes that have high connectivity with module and selected phenotype were selected as candidate genes in hub module (cor.geneModuleMembership > 0.8 and cor.geneTraitSignificance > 0.2). Then the protein/gene interactions for candidate genes were analyzed using STRING (Szklarczyk et al., 2019) and the genes connected with more than five nodes in PPI network were selected as hub genes for further study. As for selecting hub miRNAs, two web tools, microT-CDS[3] and TargetScan[4], were employed to predict candidate miRNAs for hub genes (Paraskevopoulou et al., 2013; Agarwal et al., 2015), and the score of microT-CDS > 0.9 and context $+ +$ score of TargetScan > 0.4 were selected as threshold. Then the common candidate miRNAs in hub module and prediction by microT-CDS and TargetScan were defined as real hub miRNAs. To verify our results, GSE10667 (including 15 normal lung tissues and 31 IPF tissues) and GSE70866 (including 20 normal lung tissues and 110 IPF tissues), were used to validate the different expression levels of hub genes between normal tissue and IPF tissues with two-tailed student's *t*-tests, separately.

Gene Set Enrichment Analysis (GSEA) and Guilt of Association for Hub Genes

Gene set enrichment analysis (GSEA) analysis was performed for hub genes in GSE32537 (Subramanian et al., 2005). In GSE32537, according to the median expression of this hub gene, 119 cases

[3]http://www.microrna.gr/microT-CDS/

[4]http://www.targetscan.org/

were classified into high expression group and low expression group (high group, $n = 60$; low group, $n = 59$). $|ES| > 0.5$, nominal $P < 0.05$, and FDR $\geq 25\%$ were chosen as the cut-off criteria. Besides, Spearman correlation analysis was performed to explore pair-wise gene expression correlation for hub genes in GSE10667. We calculated correlation coefficient absolute values, and the top 300 genes of each hub gene were selected for functional enrichment analysis. Based on the results, the potential functions of each hub gene were predicted, and the method thus bore the name of "guilt of association."

Construction of Hub miRNA and Hub Gene Interaction Network

According to the score of microT-CDS and the context + + score of TargetScan, miRNA–gene interaction network was constructed in Cytoscape (Shannon et al., 2003). And the interaction between genes was also demonstrated from STRING. Furthermore, text mining of hub genes and hub miRNAs was performed using GenCLip 2.0[5]. GenCLip 2.0 is an online text-mining server, which can provide the analysis of gene and miRNA functions with free terms generated by literature mining (Wang et al., 2014).

Analysis of the Role of the Interaction Network Involved in IPF and NSCLC

To further understand the role of hub genes and hub miRNAs in clinical practice, we selected two data sets (GSE70866 and GSE27430) with clearer clinical information to do clinicopathological correlation analysis separately. From GSE70866, 110 samples with IPF were used to determine the association between age and hub genes expression levels, between gender and hub genes expression levels by Pearson Chi-square test. From GSE27430, 13 samples with IPF were used to determine the association between age and hub miRNAs expression levels, gender, and hub miRNAs expression levels with Fisher test due to small sample size. P-value < 0.05 was considered as statistical significance. In addition, to explore the role of the interaction network in NSCLC (mainly including LUAD and LUSC), UALCAN[6] was used to explore the different expression levels of hub genes between normal tissues and cancer tissues (including LUAD and LUSC), separately. UALCAN is a useful online tool for analyzing cancer transcriptome data, which is based on public cancer transcriptome data (TCGA and MET500 transcriptome sequencing) (Chandrashekar et al., 2017). Moreover, we evaluate the relationship between the expression levels of hub genes and the prognosis of LUAD and LUSC, the expression levels of hub miRNAs and the prognosis of LUAD and LUSC. Kaplan Meier Plotter[7], including the gene expression data and survival information of GEO and TCGA repositories, was used to explore the relationship between the expression levels of hub genes and the survival time of LUAD and LUSC (Gyoerffy et al., 2014). Besides, OncoLnc[8], containing survival data from 21 cancer studies performed by TCGA and giving users the ability to create publication-quality Kaplan–Meier plots, was used to explore the relationship between the expression levels of hub miRNAs and the survival time of LUAD and LUSC (Anaya, 2016).

[5]http://ci.smu.edu.cn/

[6]http://ualcan.path.uab.edu/

[7]http://kmplot.com/analysis/

[8]http://www.oncolnc.org/

RESULTS

Weighted Co-expression Networks Construction and Key Modules Identification

It is found that the median of miRNA/gene expression value of each sample is approximately equal (**Supplementary Figure S1**), and the results indicated that the processed datasets can be used for further analysis. With the method of average linkage hierarchical clustering, the samples of both data sets (GSE32537 and GSE32538) are well clustered separately. The clustering dendrograms of the genes of GSE32537 are generated in **Figure 2A**, while miRNAs of GSE32538 are shown in **Figure 2B**. By "WGCNA" package in R, the genes and the miRNAs which had similar expression levels were divided into modules to construct co-expression networks. Power of $\beta = 3$ (scale free $R^2 = 0.92$) was selected as the soft-thresholding parameter for gene co-expression networks (**Supplementary Figure S2**), and power of $\beta = 5$ (scale free $R^2 = 0.89$) was selected for miRNA co-expression networks (**Supplementary Figure S3**). In gene co-expression networks, 11 modules were identified and blue module (GS = 0.38, p-value = 6.8e-282) showed the highest correlation with the scores of SGRQ. In miRNA co-expression networks, five modules were identified and turquoise module (GS = 0.20, p-value = 7.9e-58) showed the highest correlation with the scores of SGRQ (**Figure 3**). There are 285 genes in blue module and 163 miRNAs in turquoise module. Blue module (G blue) and turquoise module (M turquoise) were picked for following analysis as the clinically significant module.

Pathway Enrichment Analysis of Genes and miRNAs in Hub Modules

To explore the biological functions of the G blue, the genes were categorized into BP, CC, and MF. The outcome of GO and KEGG enrichment of the genes in blue module was shown in **Figure 4A**. The genes in BP were generally enriched in cell adhesion, extracellular matrix organization, signal transduction, positive regulation of cell proliferation, and negative regulation of cell proliferation; the genes in CC were mainly focused on plasma membrane, extracellular region, extracellular space, extracellular exosome, and extracellular matrix; the genes in MF were significantly focused on calcium ion binding, heparin binding, integrin binding, extracellular matrix structural constituent, and growth factor activity. The top five significantly enriched pathways in blue module were PI3K-Akt signaling pathway, focal adhesion, pathways in cancer, ECM–receptor interaction, and protein digestion and absorption. Top enriched GO terms for the miRNAs in turquoise module were: BP, transport, response to stress, cell death, and cell proliferation in BP; organelle, protein complex, cytosol, CC, and focal adhesion in CC; ion binding, MF, enzyme binding, RNA binding, and

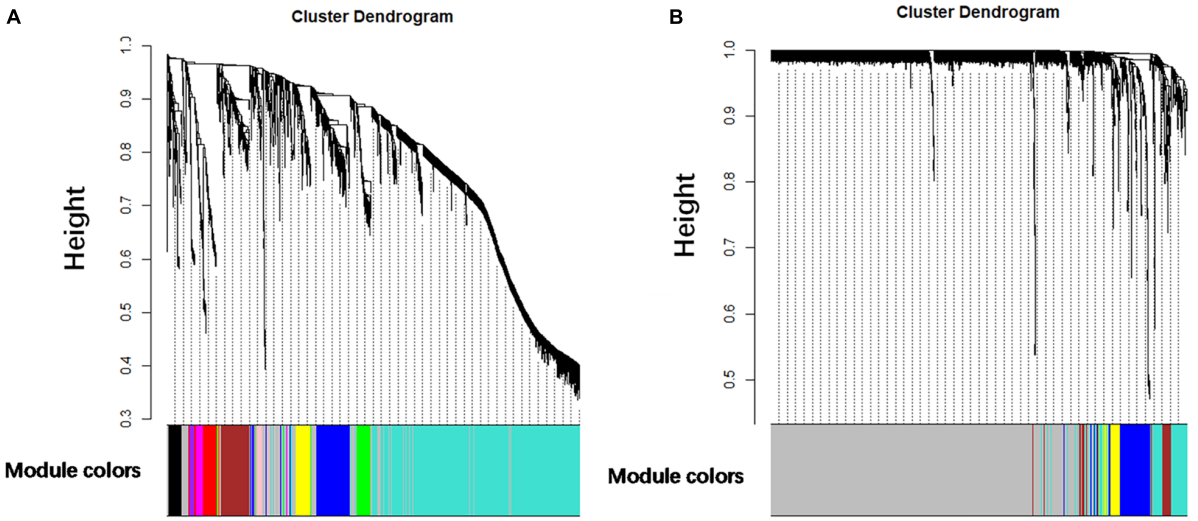


FIGURE 2 | Clustering dendrograms. **(A)** Clustering dendrograms of genes based on a dissimilarity measure (1-TOM). **(B)** Clustering dendrograms of miRNAs based on a dissimilarity measure (1-TOM).

protein binding transcription factor activity in MF. The pathway analysis was also performed for the miRNAs in turquoise module. The top five significantly enriched pathways were proteoglycans in cancer, protein processing in endoplasmic reticulum, viral carcinogenesis, pathways in cancer, and focal adhesion (**Figure 4B**).

Identification and Validation of Hub Genes and miRNAs in IPF

Under the threshold of $|MM| > 0.8$ and $|GS| > 0.2$, 58 genes in blue module were considered as candidate genes. Then the relationship between candidate genes was identified from STRING (**Supplementary Figure S4**), and we calculated the connectivity degree of each node in PPI. The nodes with degrees $\geqq 5$ were COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5, which were considered as real hub gens because it interacted with more proteins. Based on the prediction of microT-CDS and TargetScan, seven hub miRNAs (hsa-let-7b-5p, hsa-miR-26a-5p, hsa-miR-25-3p, hsa-miR-29c-3p, hsa-let-7c-5p, hsa-miR-29b-3p, and hsa-miR-26b-5p) were identified in turquoise module. In the blue module, COL3A1 and COL1A2 were the most central genes with the degrees of 13, and they are involved in the process of other genes regulating cell metabolism. As for the miRNAs, hsa-let-7b-5p was considered as key miRNA with the highest MM (MM = 0.915). The corresponding MM and GS of hub genes and hub miRNAs are shown in **Table 1**. From the results of two-tailed student's *t*-tests for GSE10667 and GSE70866, the expression levels of all hub genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) were significantly higher in IPF tissues (**Figure 5**). And the ROC curve analysis for GSE10067 indicated that the hub genes exhibited excellent diagnostic efficiency for normal tissues and IPF tissues (**Supplementary Figure S5**).

GSEA and Guilt of Association

Gene set enrichment analysis was performed to identify the lurking mechanisms related to IPF progression of six hub genes. As shown in **Supplementary Table S2**, IPF samples in COL3A1 high expression group were most significantly enriched in cellular adhesion molecules; IPF samples in COL1A2, OGN, COL15A1, ASPN, and MXRA5 high expression groups were most significantly enriched in ECM receptor interaction (**Supplementary Tables S2–S7**). Based on the analysis of guilt of association, we identified that the hub genes were essential for extracellular environment and ossification, and they mainly played important roles in extracellular structure organization, extracellular matrix

TABLE 1 | The hub genes and hub miRNAs as well as the corresponding MM and GS.

	Symbol	Degrees in PPI	MM	GS
Hub genes	COL3A1	13	0.812933	0.582487
	COL1A2	13	0.821623	0.555745
	OGN	6	0.862299	0.475489
	COL15A1	5	0.860161	0.600621
	ASPN	5	0.854841	0.642866
	MXRA5	5	0.805921	0.592231
Hub miRNAs	hsa-let-7b-5p	–	0.915297	−0.35161
	hsa-miR-26a-5p	–	0.825955	−0.44743
	hsa-miR-25-3p	–	0.793815	−0.31258
	hsa-miR-29c-3p	–	0.676672	−0.25468
	hsa-let-7c-5p	–	0.660136	−0.1454
	hsa-miR-29b-3p	–	0.622602	−0.19913
	hsa-miR-26b-5p		0.577243	−0.13157

miRNAs: microRNAs. PPI: protein/gene interactions. MM: cor.geneModuleMembership. GS: cor.geneTraitSignificance.

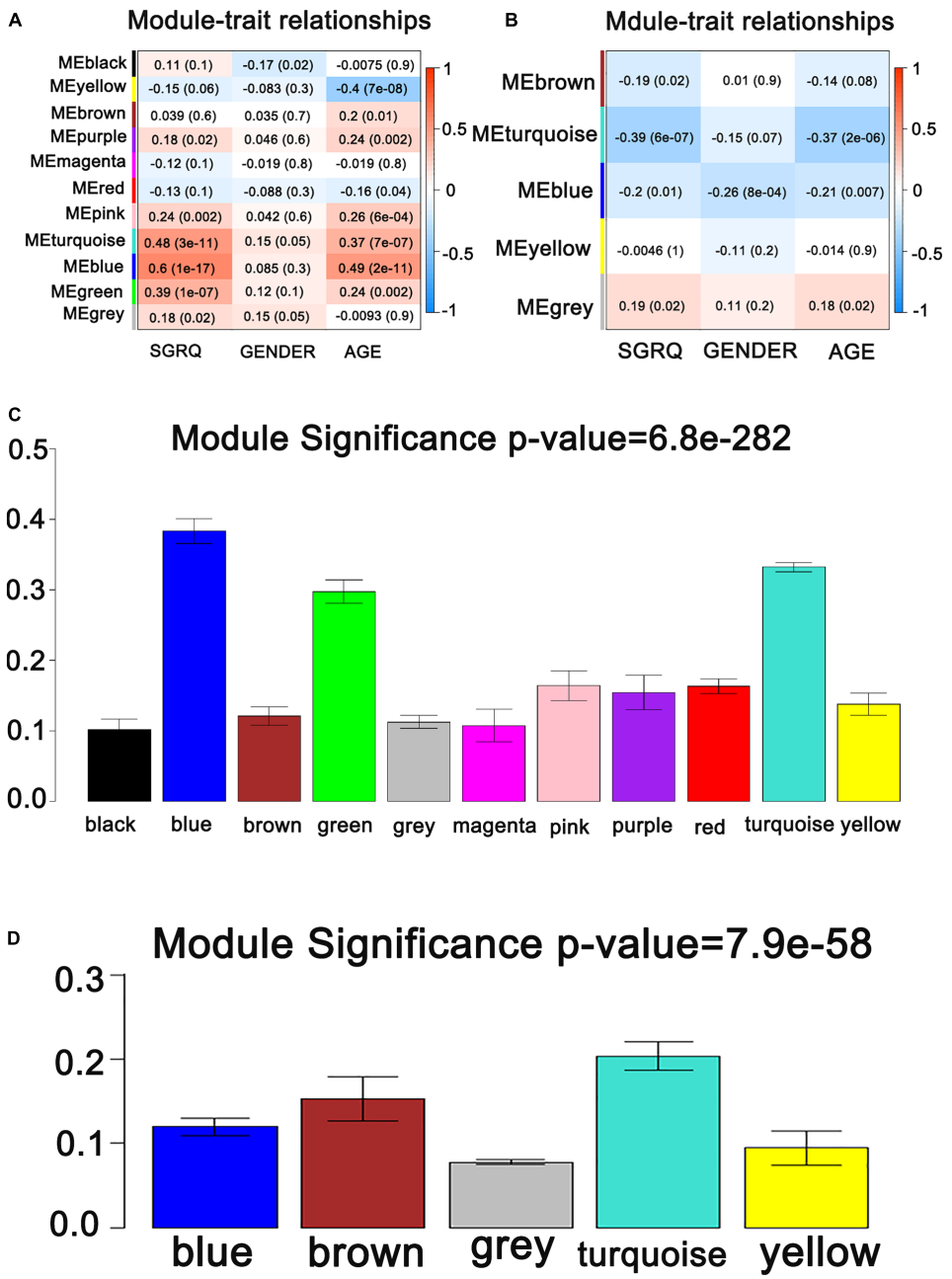


FIGURE 3 | Identification of modules associated with the clinical traits of IPF. **(A)** Heatmap of the correlation between co-expressed gene module eigengenes and clinical traits of IPF. **(B)** Heatmap of the correlation between co-expressed miRNA module eigengenes and clinical traits of IPF. **(C)** Distribution of average gene significance and errors in the modules associated with the scores of SGRQ. **(D)** Distribution of average miRNA significance and errors in the modules associated with the scores of SGRQ in IPF.

organization, and skeletal system development (**Supplementary Figure S6**).

Construction of Hub miRNA and Hub Gene Interaction Network

The hub genes and hub miRNAs interactions were predicted by microT-CDS and Targetscan (**Table 2**), and the hub genes and hub miRNAs interaction network was shown in **Figure 6A**. Six genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) and seven miRNAs (hsa-let-7b-5p, hsa-miR-26a-5p, hsa-miR-25-3p, hsa-miR-29c-3p, hsa-let-7c-5p, hsa-miR-29b-3p, and hsa-miR-26b-5p) were involved in this interaction network. Besides, the occurrence frequency of terms of corresponding literature was demonstrated from GenCLip 2.0, including extracellular matrix, transforming growth factor, squamous

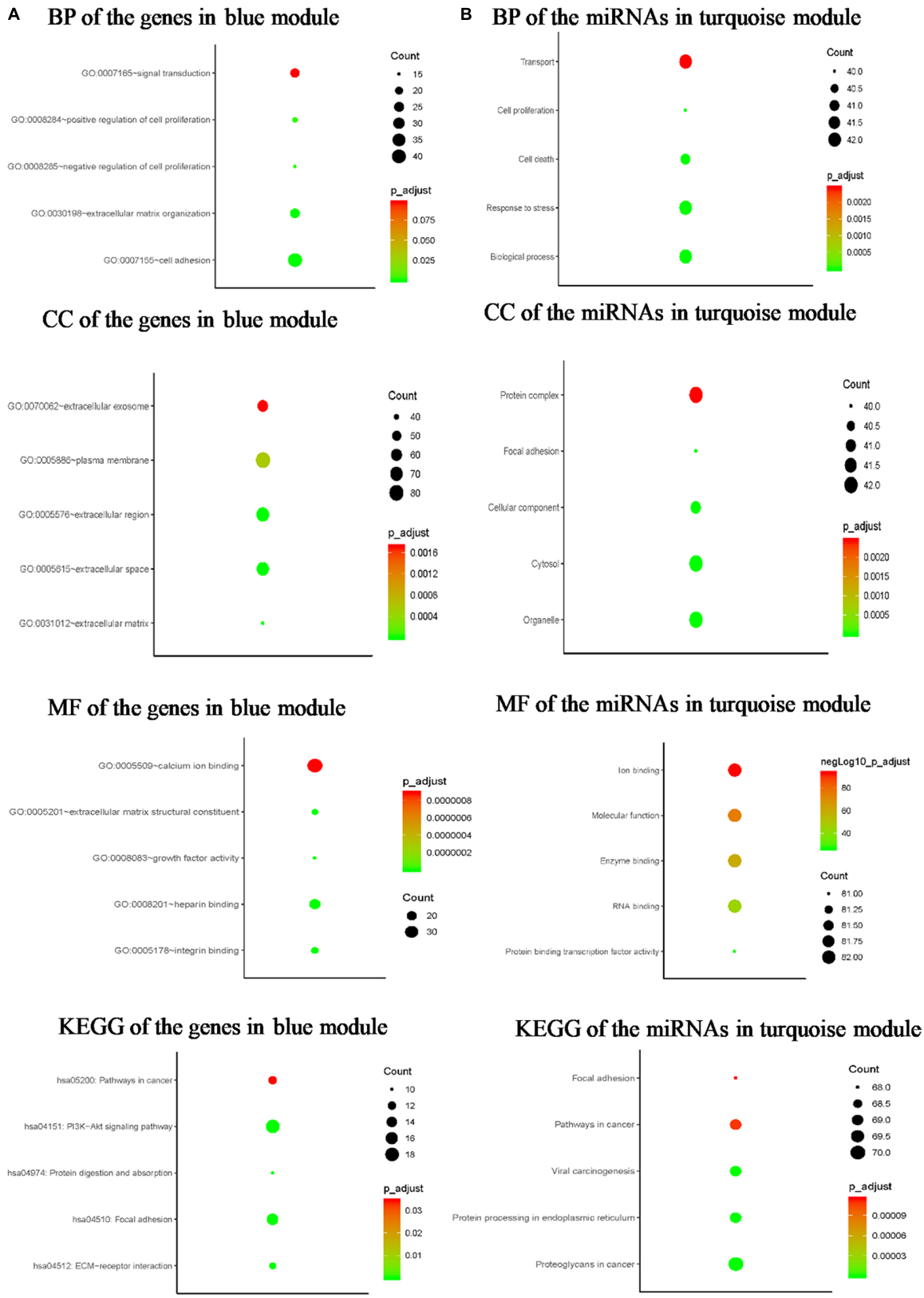


FIGURE 4 | Bioinformatics analysis of the genes in blue module and the miRNAs in turquoise module. **(A)** GO analysis and KEGG pathway enrichment of the genes in blue module. **(B)** GO analysis and KEGG pathway enrichment of the miRNAs in turquoise module.

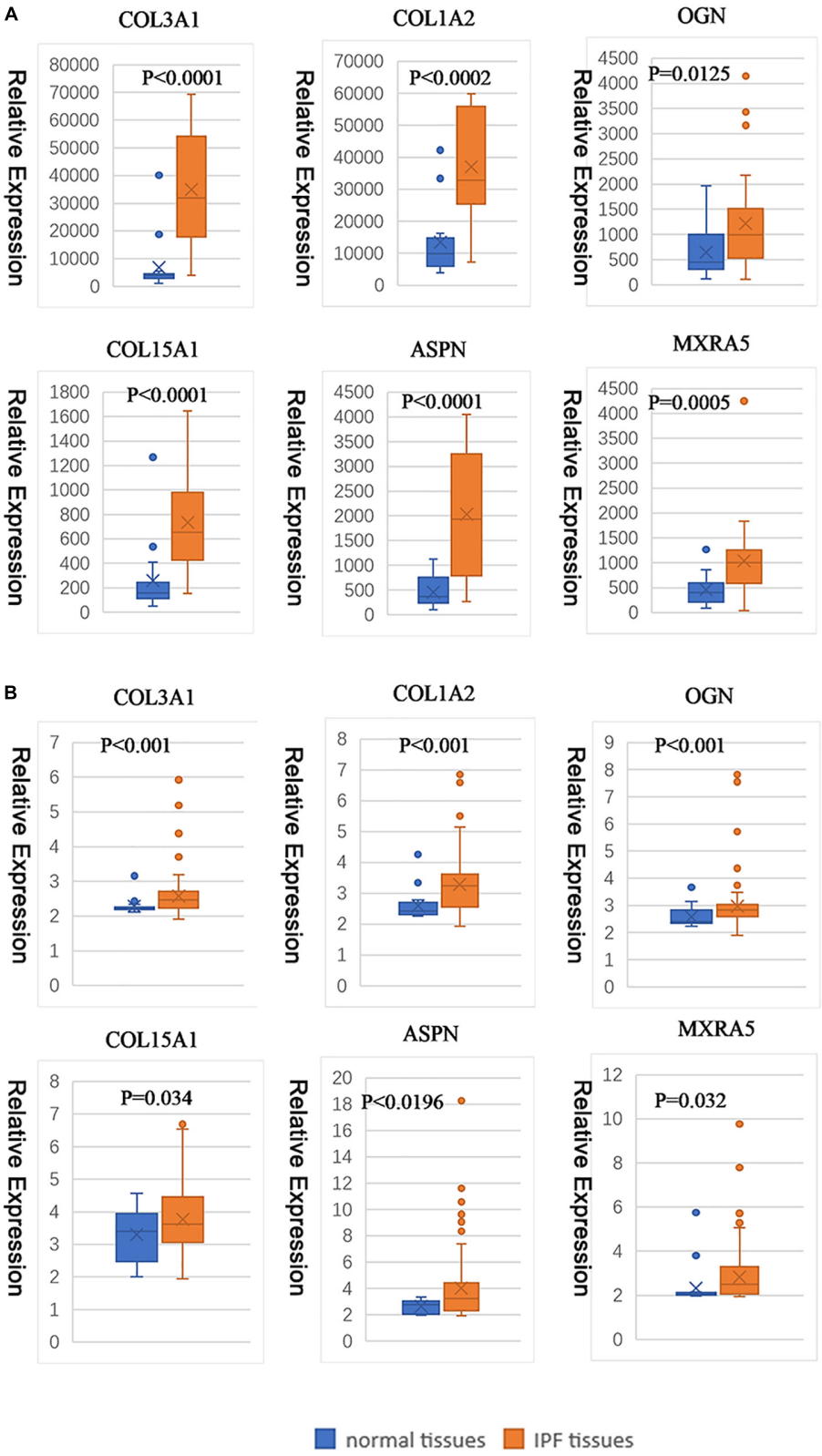


FIGURE 5 | Hub gene expression levels between normal tissue and IPF tissue (based on GSE10667 and GSE70866). The gene expression levels of COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 in GSE10667 **(A)**. The gene expression levels of COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 in GSE70866 **(B)**.

TABLE 2 | The prediction of the interaction of hub genes and hub miRNAs by microT-CDS and Targetscan.

miRNA	Target gene	Score of microT-CDS	Context + + score of TargetScan
hsa-let-7b-5p	COL3A1	0.99	−0.47
hsa-let-7b-5p	COL1A2	0.99	−0.5
hsa-miR-26a-5p	ASPN	0.99	−0.41
hsa-miR-25-3p	ASPN	0.93	−0.44
hsa-miR-29c-3p	COL3A1	0.99	−0.87
hsa-miR-29c-3p	COL1A2	0.99	−0.61
hsa-miR-29c-3p	COL15A1	0.99	−0.5
hsa-let-7c-5p	COL3A1	0.99	−0.47
hsa-let-7c-5p	COL1A2	0.99	−0.5
hsa-miR-29b-3p	COL3A1	0.99	−0.87
hsa-miR-29b-3p	COL1A2	0.99	−0.61
hsa-miR-29b-3p	COL15A1	0.99	−0.52
hsa-miR-26b-5p	ASPN	0.98	−0.4

miRNA: microRNA.

cell carcinoma, mesenchymal stem cell, fibrillar collagen, procollagen, and osteoblast differentiation (**Figure 6B**).

Analysis of Hub Genes–Hub miRNAs Interaction Network in IPF and NSCLC

Based on the results of clinicopathological correlation analysis, there were no statistical differences in age distribution and gender distribution between these high-expression and low-expression groups of hub genes. And we also did not find any substantial differences in age distribution and gender distribution between these high-expression and low-expression groups of hub miRNAs. More details are listed in **Supplementary Table S8**. Furthermore, some databases were used to explore the role of the interaction network in NSCLC (LUAD and LUSC). The levels of the six genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) expression were significantly different between normal samples and LUAD samples from UALCAN (**Figure 7A**). COL3A1, COL1A2, COL15A1, ASPN, and MXRA5 were higher expressed in tumor samples, while OGN was lower expressed. In LUSC tissues, the levels of COL3A1, COL1A2, OGN, ASPN, and MXRA5 expressions were significantly different from normal lung tissues, and there is no difference of COL15A1 between normal tissues and LUSC tissues (**Figure 7B**). For the relationship between hub genes expression levels and the prognosis of NSCLC from Kaplan Meier Plotter, COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 were associated with the overall survival of LUAD (**Figure 8A**), but the expression levels of these genes did not affect overall survival of LUSC patients. Besides, hsa-miR-29c-3p, hsa-let-7c-5p, hsa-miR-29b-3p were identified to be related to the overall survival of LUAD from OncoLnc (**Figure 8B**).

DISSCUSION

Idiopathic pulmonary fibrosis is a medically incurable disease with complicated clinical manifestations. Nowadays, only two medicines, nintedanib and pirfenidone, are approved for the treatment to slow down the progression of IPF (Lehtonen et al., 2016; Maher et al., 2017; Drakopanagiotakis et al., 2018). In order to identify a meaningful biomarker, a part of previous studies had focused too much on single miRNA or gene (Mizuno et al., 2017), and this cannot meet the requirement for exploration of molecular mechanisms in IPF progression. Though another part of previous studies had reported differently expressed genes and differently expressed miRNAs between normal tissue and IPF tissues to further explore the molecular mechanisms, the relationships between hubs and important clinical traits had not been rigorously studied, which would make clinically significance few. Besides, there are some previous studies focusing preclinical models by aberrant gene expression; though these modules are useful for clinical application, it did not make much sense in exploration of pathogenesis in IPF and NSCLC. It is a pity that the research on molecular mechanisms of IPF affecting NSCLC occurrence and prognosis was little, especially in bioinformatics. To fulfill these gaps, the interaction network of hub miRNAs-hub genes was studied on this research, and WGCNA was used to identify IPF gene and miRNA modules for the first time. More importantly, it was the first time to explore the common mechanisms and molecular targets between IPF and NSCLC in bioinformatics, which would provide more information about that IPF causing NSCLC and poor NSCLC prognosis, and this more attention is to be called on IPF-NSCLC patients. Two modules were found, including one gene module (blue module) and one miRNA module (turquoise module), were significantly related to the scores of SGRQ. We identified six hub genes and seven hub miRNAs, and the hub miRNAs–hub genes interaction network was constructed. In GenCLip 2.0, the BPs (extracellular matrix, transforming growth factor, squamous cell carcinoma, mesenchymal stem cell, etc.) were considered to be significantly related to IPF and NSCLC.

Focal adhesion was considered as a key pathway shared by blue module and turquoise module, and many gens/proteins have been considered to be involved in the progression of IPF through disordering focal adhesion (Gimenez et al., 2017; Kathiriya et al., 2017; Molina-Molina et al., 2018). For example, it has been reported that decreased expression of collagen VI, an important kind of protein of ECM, would upregulate the focal adhesion (Knueppel et al., 2018). For example, COL1A2, which is a subtype of Type I collagen (Fang et al., 2019), is implicated in the induction of epithelial–mesenchymal transition in many fibroblasts (Cheng et al., 2017). Type I collagen could induce the disruption of E−cadherin33 and SMADS to downregulate E−cadherin (Koenig et al., 2006). Of course, there are still potential pathways worth further study about hub genes in IPF. In present study, the hub miRNAs, except hsa-miR-25-3p (Min et al., 2016), were identified to be related to the progression of IPF for the first time, which would be novel diagnostic biomarkers of patients with IPF.

After analyzing and comparing the results of GSEA analysis and guilt of association, we found that ECM–receptor interaction is an important pathway shared by hub genes. Pulmonary extracellular matrix, which is a complex system composed of proteoglycans and glycosaminoglycans, is of importance in

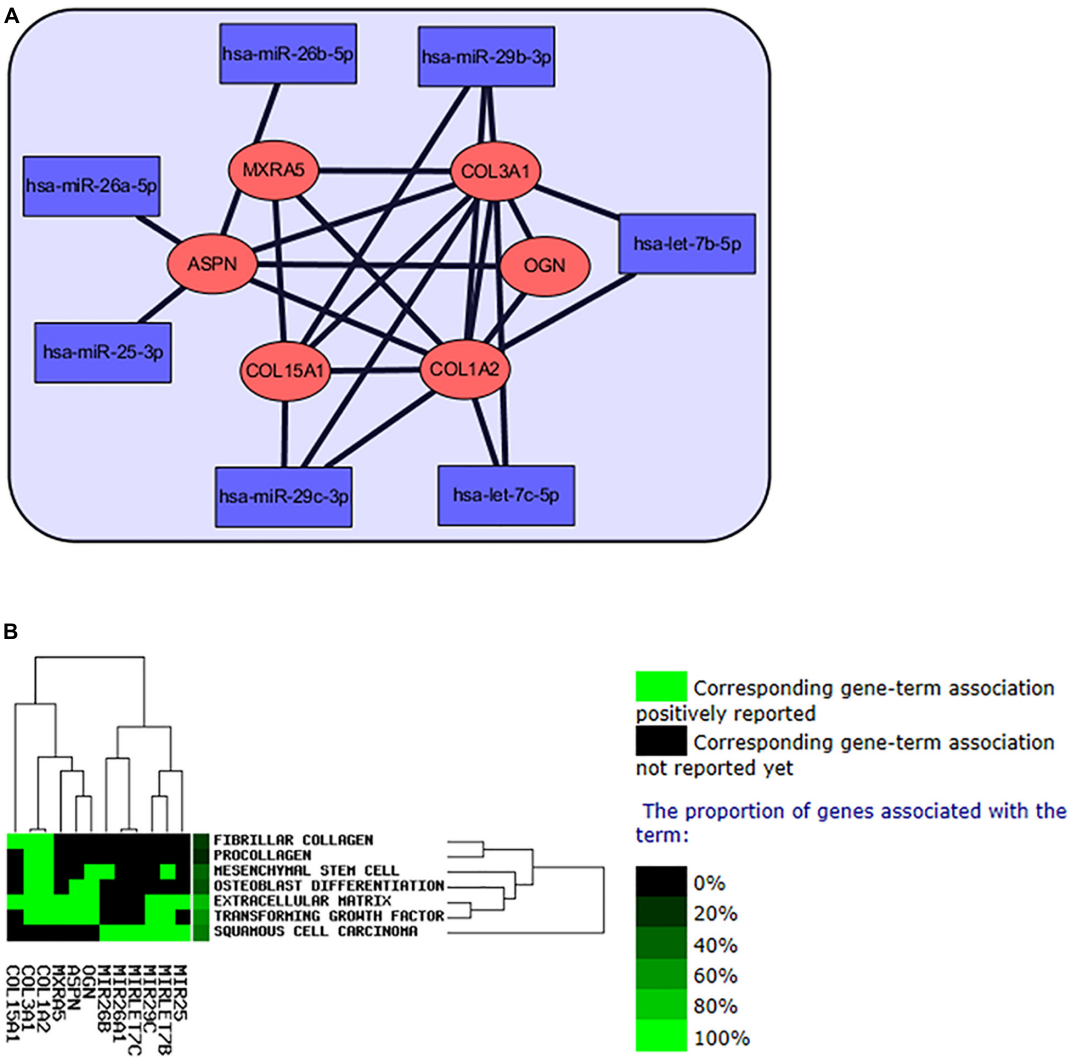


FIGURE 6 | The interaction network of hub miRNAs and hub genes. **(A)** The network of regulation of hub miRNAs and hub genes in IPF. **(B)** Text mining of the hub genes and hub miRNAs from GenCLip 2.0 software.

tissue's homeostasis and repair. Previous studies have revealed that ECM protein expression plays an important role in the fibrotic process in IPF lungs (Vicens-Zygmunt et al., 2015). Excessive accumulation of ECM in the alveolar parenchyma and progressive scarring of lung tissue are major characteristics of IPF (Knudsen et al., 2017), and some studies have used this protein expression level as a criterion for evaluating treatment outcomes (Molina-Molina et al., 2018; Mullenbrock et al., 2018). Altogether, migration is strongly influenced by topology and composition of the ECM including integrin ligands, and the hub gens and hub miRNAs might play an important role in IPF progression with the change of ECM.

Evidence suggests that patients with NSCLC who develop IPF have worse outcomes than patients without IPF (Han et al., 2019). Clinical examples with both diseases are numerous, and they are difficult to treat. In the treatment of patients suffered IPF and NSCLC, physicians are reluctant to treat NSCLC because of the poor prognosis of IPF (Kinoshita and Goto, 2019). Therefore, the interaction network was analyzed between these two types of diseases, which would provide more information about that IPF causing NSCLC and poor NSCLC prognosis. Though cancer was not taken as the main research topic at first, with analysis continuing, we identified hub miRNAs and hub genes may participate in the progression of NSCLC. And the hub miRNAs–hub genes interaction network would help us understand the pathogenesis of IPF-NSCLC. For example, COL3A1 is highly expressed in both IPF and NSCLC tissues, so it is speculated that COL3A1 is a key molecule of cross-linking between IPF and NSCLC, and even a signal of IPF leading to NSCLC. MXRA5 is upregulated in IPF, and it is found that the higher the expression, the worse the prognosis of NSCLC. We speculated that MXRA5 is an important intermediate molecule of IPF leading to poor prognosis of NSCLC. Of course, these all need further experimental verification later, and some experiments

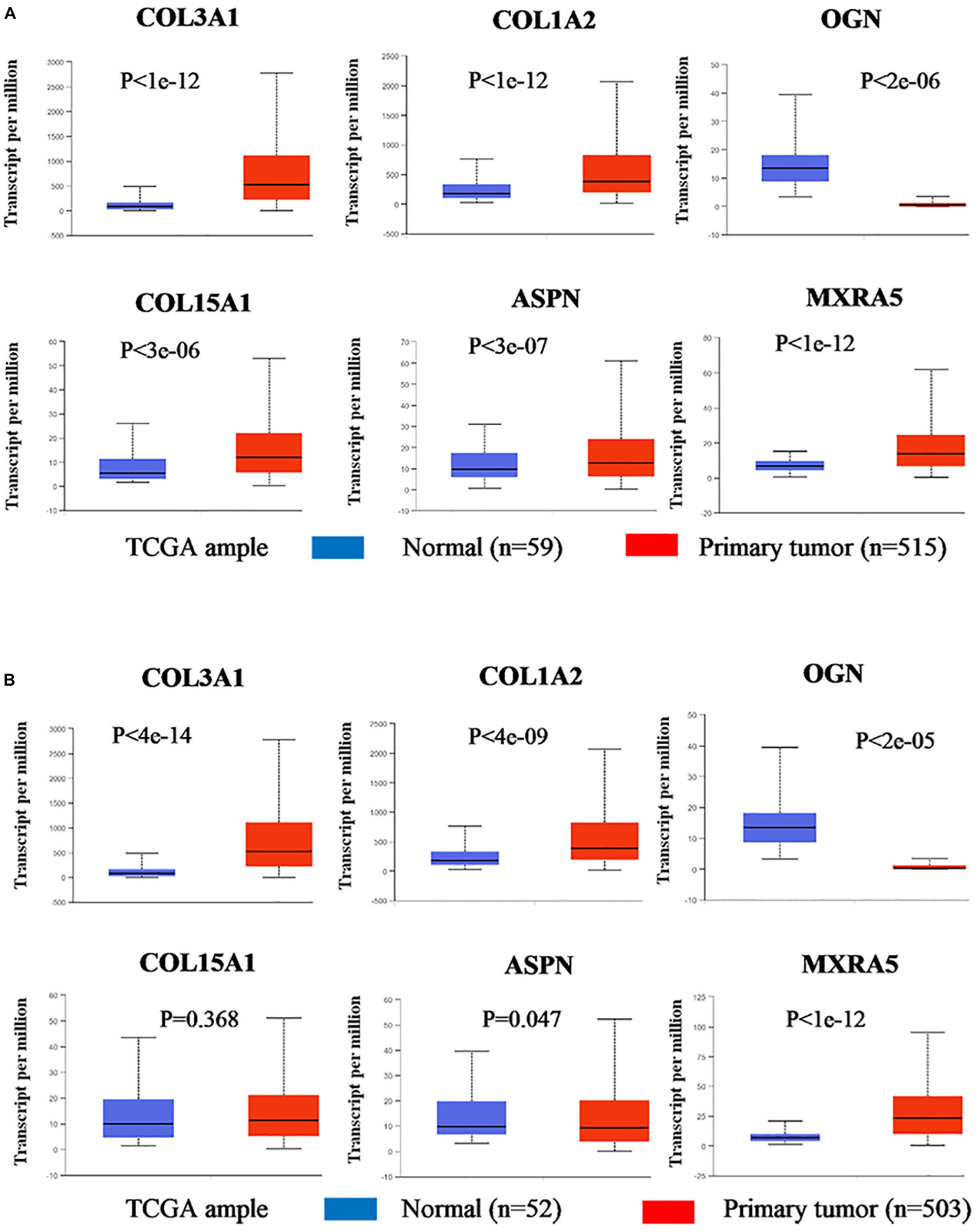


FIGURE 7 | Gene expression levels between normal lung and tumor samples (based on TCGA data in UALCAN). **(A)** The gene expression levels of COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 between normal lung tissues and LUAD tissues. **(B)** The gene expression levels of COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 between normal lung tissues and LUSC tissues.

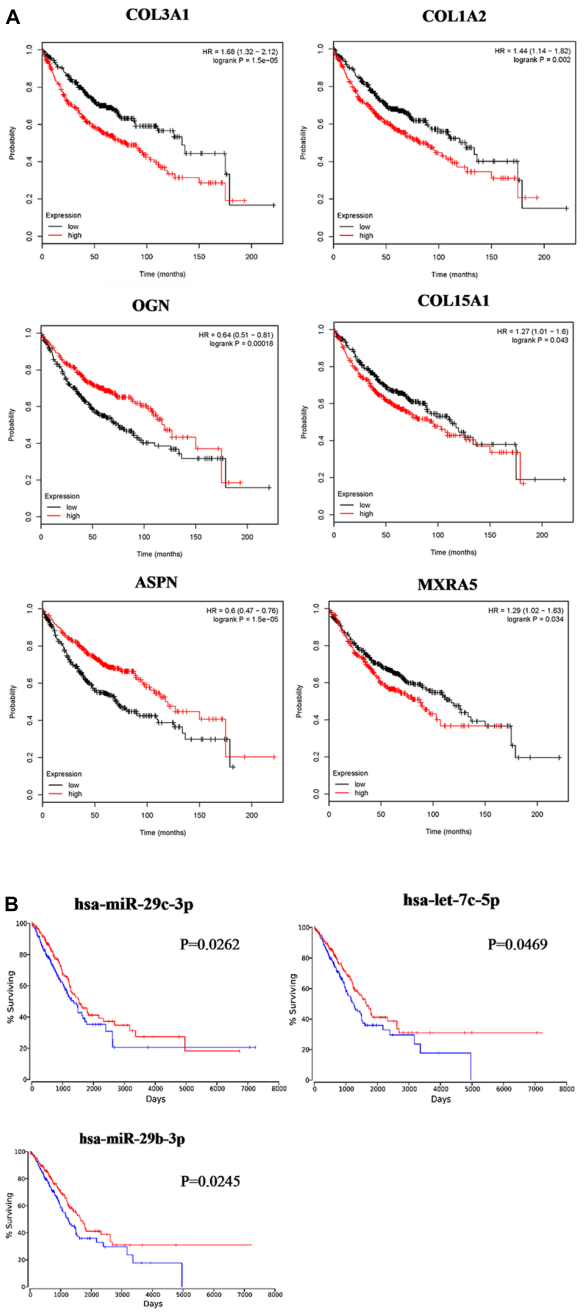


FIGURE 8 | Survival analysis of the association between the expression levels of hub genes and hub miRNAs in LUAD. COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 were identified to be related to the overall survival of LUAD from Kaplan Meier Plotter **(A)**. hsa-miR-29c-3p, hsa-let-7c-5p, and hsa-miR-29b-3p were identified to be related to the overall survival of LUAD from OncoLnc **(B)**.

need to be done to confirm the hub genes. We will further explore the hubs and its role in the progression of IPF-NSCLC by using more in-depth bioinformatic analyses and experimental methods in the future., In this study, OGN was identified to be related to the progression of IPF for the first time. Most interestingly, we found that OGN is highly expressed in IPF, but is lowly expressed in cancer tissues. And low expression levels of OGN would have an important impact on the prognosis of LUAD (**Figure 8**). Different signal pathways should be activated to regulate or influence OGN. Although many studies identified that the expression levels of OGN would alter in cancers, such as gastric cancer (Lee et al., 2003), colorectal cancer (Hu et al., 2018), and invasive ductal breast carcinoma (Roewer et al., 2011), functional data about how OGN participating in cancer pathology are not enough, and further studies are needed.

CONCLUSION

It was the first time to construct miRNA–gene interaction network to explore the development of IPF and common pathways between IPF and NSCLC by WGCNA. We identified six hub genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5) and seven hub miRNAs (hsa-let-7b-5p, hsa-miR-26a-5p, hsa-miR-25-3p, hsa-miR-29c-3p, hsa-let-7c-5p, hsa-miR-29b-3p, and hsa-miR-26b-5p), which might be diagnostic biomarkers for IPF. In the future, the pathogenic overlap of IPF and NSCLC may help us to clarify the common molecular mechanisms between both diseases, and may provide a potential treatment strategy for both diseases.

DATA AVAILABILITY STATEMENT

The data analyzed in this study can be found in the GEO database (http://www.ncbi.nlmNih.gov/geo), using accession numbers GSE32537, GSE10667, GSE70866, GSE32538, and GSE27430.

AUTHOR CONTRIBUTIONS

DY, X-LR, X-PL, and SL reviewed relevant literature and drafted the manuscript. DY, X-LR, J-YH, H-LM, CC, and W-DH conducted all statistical analyses. All authors read and approved the final manuscript.

FUNDING

This work was supported by the 351 Talent Project of Wuhan University (Luojia Young Scholars: SL) and Young & Middle-aged Medical Key Talents Training Project of Wuhan (WHQG201901).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00302/full#supplementary-material

FIGURE S1 | Standardization of gene expression. The data quality was evaluated, and boxplot was used to compare before and after being standardized.

FIGURE S2 | Determination of soft-thresholding power in the weighted gene co-expression network analysis (WGCNA). **(a)** Analysis of the scale-free fit index for various soft-thresholding powers. **(b)** Analysis of the mean connectivity for various soft-thresholding powers. **(c)** Histogram of connectivity distribution when β = 3. **(d)** Checking the scale free topology when β = 3.

FIGURE S3 | Determination of soft-thresholding power in the weighted miRNA co-expression network analysis. **(a)** Analysis of the scale-free fit index for various soft-thresholding powers. **(b)** Analysis of the mean connectivity for various soft-thresholding powers. **(c)** Histogram of connectivity distribution when β = 5. **(d)** Checking the scale free topology when β = 5.

FIGURE S4 | Protein–protein interaction network of 58 candidate genes acquired from STRING 9.1.

FIGURE S5 | ROC curve of COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5 in GSE10067.

FIGURE S6 | Guilt of association for hub genes (COL3A1, COL1A2, OGN, COL15A1, ASPN, and MXRA5).

TABLE S1 | Gene and miRNA expression microarray datasets related to IPF.

TABLE S2 | Gene set enriched in lung samples with COL3A1 high expression.

TABLE S3 | Gene set enriched in lung samples with COL1A2 high expression.

TABLE S4 | Gene set enriched in lung samples with OGN high expression.

TABLE S5 | Gene set enriched in lung samples with COL15A1 high expression.

TABLE S6 | Gene set enriched in lung samples with ASPN high expression.

TABLE S7 | Gene set enriched in lung samples with MXRA5 high expression.

TABLE S8 | Clinicopathological correlation analysis for hub genes and hub miRNAs in IPF.

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005

Anaya, J. (2016). OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peerj Comput. Sci.* 2:e67. doi: 10.7717/peerj-cs.67

Ballester, B., Milara, J., and Cortijo, J. (2019). Idiopathic pulmonary fibrosis and lung cancer: mechanisms and molecular targets. *Int. J. Mol. Sci.* 20:593. doi: 10.3390/ijms20030593

Berschneider, B., Ellwanger, D. C., Baarsma, H. A., Thiel, C., Shimbori, C., White, E. S., et al. (2014). miR-92a regulates TGF-beta 1-induced WISP1 expression in pulmonary fibrosis. *Int. J. Biochem. Cell Biol.* 53, 432–441. doi: 10.1016/j.biocel.2014.06.011

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Cheng, Z., Gao, W., Fan, X., Chen, X., Mei, H., Liu, J., et al. (2017). Extracellular signal-regulated kinase 5 associates with casein kinase II to regulate GPIb-IX-mediated platelet activation via the PTEN/PI3K/Akt pathway. *J. Thromb. Haemost.* 15, 1679–1688. doi: 10.1111/jth.13755

Datta, A., Scotton, C. J., and Chambers, R. C. (2011). Novel therapeutic approaches for pulmonary fibrosis. *Br. J. Pharmacol.* 163, 141–172. doi: 10.1111/j.1476-5381.2011.01247.x

Drakopanagiotakis, F., Wujak, L., Wygrecka, M., and Markart, P. (2018). Biomarkers in idiopathic pulmonary fibrosis. *Matrix Biol.* 68-69, 404–421. doi: 10.1016/j.matbio.2018.01.023

Fan, L., Yu, X., Huang, Z., Zheng, S., Zhou, Y., Lv, H., et al. (2017). Analysis of microarray-identified genes and micrornas associated with idiopathic pulmonary fibrosis. *Mediators Inflammation* 2017:9. doi: 10.1155/2017/1804240

Fang, S., Dai, Y., Mei, Y., Yang, M., Hu, L., Yang, H., et al. (2019). Clinical significance and biological role of cancer-derived Type I collagen in lung and esophageal cancers. *Thoracic Cancer* 10, 277–288. doi: 10.1111/1759-7714.12947

Gao, S.-Y., Zhou, X., Li, Y.-J., Liu, W.-L., Wang, P.-Y., Pang, M., et al. (2014). Arsenic trioxide prevents rat pulmonary fibrosis via miR-98 overexpression. *Life Sci.* 114, 20–28. doi: 10.1016/j.lfs.2014.07.037

Gimenez, A., Duch, P., Puig, M., Gabasa, M., Xaubet, A., and Alcaraz, J. (2017). Dysregulated collagen homeostasis by matrix stiffening and tgf-1 in fibroblasts from idiopathic pulmonary fibrosis patients: role of FAK/Akt. *Int. J. Mol. Sci.* 18:2431. doi: 10.3390/ijms18112431

Gyoerffy, B., Surowiak, P., Budczies, J., and Lanczky, A. (2014). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell Lung Cancer. *PLoS One* 9:e0111842. doi: 10.1371/journal.pone.0111842

Han, S. Y., Lee, Y. J., Park, J. S., Cho, Y.-J., Yoon, H. I., Lee, J.-H., et al. (2019). Prognosis of non-small-cell lung cancer in patients with idiopathic pulmonary fibrosis. *Sci. Rep.* 9:12561. doi: 10.1038/s41598-019-49026-y

Hu, X., Li, Y.-Q., Li, Q.-G., Ma, Y.-L., Peng, J.-J., and Cai, S.-J. (2018). Osteoglycin (OGN) reverses epithelial to mesenchymal transition and invasiveness in colorectal cancer via EGFR/Akt pathway. *J. Exp. Clin. Cancer Res.* 37:41. doi: 10.1186/s13046-018-0718-2

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huleihel, L., Ben-Yehudah, A., Milosevic, J., Yu, G., Pandit, K., Sakamoto, K., et al. (2014). Let-7d microRNA affects mesenchymal phenotypic properties of lung fibroblasts. *Am. J. Physiol. Lung. Cell. Mol. Physiol.* 306, L534–L542. doi: 10.1152/ajplung.00149.2013

Kathiriya, J. J., Nakra, N., Nixon, J., Patel, P. S., Vaghasiya, V., Alhassani, A., et al. (2017). Galectin-1 inhibition attenuates profibrotic signaling in hypoxia-induced pulmonary fibrosis. *Cell Death Discovery* 3, 17010–17010. doi: 10.1038/cddiscovery.2017.10

Kato, E., Takayanagi, N., Takaku, Y., Kagiyama, N., Kanauchi, T., Ishiguro, T., et al. (2018). Incidence and predictive factors of lung cancer in patients with idiopathic pulmonary fibrosis. *ER J. Open Res.* 4, 00111–2016. doi: 10.1183/23120541.00111-2016

Kinoshita, T., and Goto, T. (2019). Molecular mechanisms of pulmonary fibrogenesis and its progression to lung cancer: a review. *Int. J. Mol. Sci.* 20:1461. doi: 10.3390/ijms20061461

Knudsen, L., Ruppert, C., and Ochs, M. (2017). Tissue remodelling in pulmonary fibrosis. *Cell Tissue Res.* 367, 607–626. doi: 10.1007/s00441-016-2543-2

Knueppel, L., Heinzelmann, K., Lindner, M., Hatz, R., Behr, J., Eickelberg, O., et al. (2018). FK506-binding protein 10 (FKBP10) regulates lung fibroblast migration via collagen VI synthesis. *Respiratory Res.* 19:1461. doi: 10.1186/s12931-018-0768-761

Koenig, A., Mueller, C., Hasel, C., Adler, G., and Menke, A. (2006). Collagen type I induces disruption of E-cadherin-mediated cell-cell contacts and promotes proliferation of pancreatic carcinoma cells. *Cancer Res.* 66, 4662–4671. doi: 10.1158/0008-5472.can-05-2804

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Lawrence, P. J., Kolsum, U., Gupta, V., Donaldson, G., Singh, R., Barker, B., et al. (2017). Characteristics and longitudinal progression of chronic obstructive pulmonary disease in GOLD B patients. *BMC Pulm. Med.* 17:42. doi: 10.1186/s12890-017-0384-388

Lee, J. Y., Eom, E. M., Kim, D. S., Ha-Lee, Y. M., and Lee, D. H. (2003). Analysis of gene expression profiles of gastric normal and cancer tissues by SAGE. *Genomics* 82, 78–85. doi: 10.1016/s0888-7543(03)00098-3

Lee, T., Park, J. Y., Lee, H. Y., Cho, Y.-J., Yoon, H. I., Lee, J. H., et al. (2014). Lung cancer in patients with idiopathic pulmonary fibrosis: clinical characteristics and impact on survival. *Respiratory Med.* 108, 1549–1555. doi: 10.1016/j.rmed.2014.07.020

Lehtonen, S. T., Veijola, A., Karvonen, H., Lappi-Blanco, E., Sormunen, R., Korpela, S., et al. (2016). Pirfenidone and nintedanib modulate properties of fibroblasts and myofibroblasts in idiopathic pulmonary fibrosis. *Respiratory Res.* 17:14. doi: 10.1186/s12931-016-0328-325

Maher, T. M., Oballa, E., Simpson, J. K., Porte, J., Habgood, A., Fahy, W. A., et al. (2017).). An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir. Med.* 5, 946–955. doi: 10.1016/S2213-2600(17)30430-7

Min, H., Fan, S., Song, S., Zhuang, Y., Li, H., Wu, Y., et al. (2016). Plasma microRNAs are associated with acute exacerbation in idiopathic pulmonary fibrosis. *Diagn. Pathol.* 11:135.

Mizuno, K., Mataki, H., Seki, N., Kumamoto, T., Kamikawaji, K., and Inoue, H. (2017). MicroRNAs in non-small cell lung cancer and idiopathic pulmonary fibrosis. *J. Hum. Genet.* 62, 57–65. doi: 10.1038/jhg.2016.98

Molina-Molina, M., Machahua-Huamani, C., Vicens-Zygmunt, V., Llatjos, R., Escobar, I., Sala-Llinas, E., et al. (2018). Anti-fibrotic effects of pirfenidone and rapamycin in primary IPF fibroblasts and human alveolar epithelial cells. *BMC Pulm. Med.* 18:624. doi: 10.1186/s12890-018-0626-624

Mullenbrock, S., Liu, F., Szak, S., Hronowski, X., Gao, B., Juhasz, P., et al. (2018). Systems Analysis of Transcriptomic and Proteomic Profiles Identifies Novel Regulation of Fibrotic Programs by miRNAs in Pulmonary Fibrosis Fibroblasts. *Genes* 9:588. doi: 10.3390/genes9120588

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., et al. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucl. Acids Res.* 41, W169–W173. doi: 10.1093/nar/gkt393

Roewer, C., Ziems, B., Radtke, A., Schmitt, O., Reimer, T., Koy, C., et al. (2011). Toponostics of invasive ductal breast carcinoma: combination of spatial protein expression imaging and quantitative proteome signature analysis. *Int. J. Clin. Exp. Pathol.* 4, 454–467.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Swigris, J. J., Esser, D., Conoscenti, C. S., and Brown, K. K. (2014). The psychometric properties of the St George's Respiratory Questionnaire (SGRQ) in patients with idiopathic pulmonary fibrosis: a literature review. *Health Q. Life Outcomes* 12, 124. doi: 10.1186/s12955-014-0124-121

Swigris, J. J., Wilson, H., Esser, D., Conoscenti, C. S., Stansen, W., Leidy, N. K., et al. (2018). Psychometric properties of the St George's respiratory questionnaire in patients with idiopathic pulmonary fibrosis: insights from the INPULSIS trials. *Bmj Open Respi. Res.* 5:e000278. doi: 10.1136/bmjresp-2018-000278

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Taganov, K. D., Boldin, M. P., and Baltimore, D. (2007). MicroRNAs and immunity: Tiny players in a big field. *Immunity* 26, 133–137. doi: 10.1016/j.immuni.2007.02.005

Tomassetti, S., Gurioli, C., Ryu, J. H., Decker, P. A., Ravaglia, C., Tantalocco, P., et al. (2015). The impact of lung cancer on survival of idiopathic pulmonary fibrosis. *Chest* 147, 157–164. doi: 10.1378/chest.14-0359

Vicens-Zygmunt, V., Estany, S., Colom, A., Montes-Worboys, A., Machahua, C., Juliana Sanabria, A., et al. (2015). Fibroblast viability and phenotypic changes within glycated stiffened three-dimensional collagen matrices. *Respir. Res.* 16:82. doi: 10.1186/s12931-015-0277-4

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., et al. (2015). DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucl. Acids Res.* 43, W460–W466. doi: 10.1093/nar/gkv403

Wang, J.-H., Zhao, L.-F., Lin, P., Su, X.-R., Chen, S.-J., Huang, L.-Q., et al. (2014). GenCLiP 2.0: a web server for functional clustering of genes and construction of molecular networks based on free terms. *Bioinformatics* 30, 2534–2536. doi: 10.1093/bioinformatics/btu241

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 April 2020
doi: 10.3389/fgene.2020.00369

CD38 Predicts Favorable Prognosis by Enhancing Immune Infiltration and Antitumor Immunity in the Epithelial Ovarian Cancer Microenvironment

Ying Zhu[1,2†], Zhigang Zhang[1,2†], Zhou Jiang[2], Yang Liu[1,2] and Jianwei Zhou[1]*

[1] Department of Gynecology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China, [2] Key Laboratory of Tumor Microenvironment and Immune Therapy of Zhejiang Province, Hangzhou, China

The identification of predictive biomarkers and novel targets to optimize immunotherapy strategies for epithelial ovarian cancer (EOC) is urgently needed. CD38 is a multifunctional glycoprotein that acts as an ectoenzyme and immune receptor. However, the underlying immunological mechanisms and prognostic value of CD38 in EOC remain unclear. CD38 gene expression in EOC was evaluated by using Gene Expression Profiling Interactive Analysis (GEPIA) and TISIDB database. The prognostic value was calculated using GEPIA and Kaplan–Meier plotter. Gene set enrichment analysis was conducted to study the roles of CD38 in the EOC microenvironment. Furthermore, the relationship between CD38 expression level and immune cell infiltration was analyzed by the Tumor Immune Estimation Resource and TISIDB. The GEPIA and TISIDB databases showed that CD38 expression in EOC was higher than that in normal tissue and was highest in the immunoreactive subtype among the four molecular types. A total of 424 cases from GEPIA revealed that high levels of CD38 were associated with longer disease-free survival [hazard ratio (HR) = 0.66, P = 0.00089] and increased overall survival rate (HR = 0.67, P = 0.0016). Kaplan–Meier plotter also confirmed the prognostic value of CD38 in EOC. Data from The Cancer Genome Atlas database demonstrated that gene signatures in many categories, such as immune response and adaptive immune response, were enriched in EOC samples with high CD38 expression. In addition, CD38 was positively correlated with immune cell infiltration, especially infiltration of activated $CD8^+$ T cells, $CD4^+$ T cells, and B cells. CD38 is positively correlated with prognosis and immune cell infiltration in the EOC microenvironment and contributes to the regulation of antitumor immunity. CD38 could be used as a prognostic biomarker and potential immunotherapy target.

Keywords: CD38, ovarian cancer, prognosis, tumor-infiltrating lymphocytes, antitumor immunity

OPEN ACCESS

Edited by:
Wan Zhu,
Stanford University, United States

Reviewed by:
Chang Gong,
Johns Hopkins University, United States
Xinyu Chen,
Stanford University, United States

****Correspondence:***
Jianwei Zhou
2195045@zju.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 October 2019*
Accepted: *25 March 2020*
Published: *30 April 2020*

Citation:
Zhu Y, Zhang Z, Jiang Z, Liu Y and Zhou J (2020) CD38 Predicts Favorable Prognosis by Enhancing Immune Infiltration and Antitumor Immunity in the Epithelial Ovarian Cancer Microenvironment. Front. Genet. 11:369. doi: 10.3389/fgene.2020.00369

INTRODUCTION

Epithelial ovarian cancer (EOC) is the seventh most common cancer and seriously threatens female health worldwide (Siegel et al., 2019). There are no typical early symptoms and feasible screening options, and the majority of ovarian cancer patients present with late or advanced disease (stages III and IV) (Bowtell et al., 2015; Menon et al., 2018). The standard curative treatments

involve cytoreductive surgery followed by platinum-based chemotherapy. Despite improvements in therapy, relapse is inevitable, and the 5-year overall survival (OS) for EOC is approximately only 45% (Lheureux et al., 2019b). Currently, multitarget immunotherapy has become one of the most promising approaches in cancer therapy. In particular, immune checkpoint blockade, with targets such as PD-1, PD-L1, and CTLA-4, has emerged as a novel therapeutic method with noteworthy results in malignant melanoma and lung cancer (Ribas and Wolchok, 2018; Scott et al., 2018). In general, immunotherapy is less efficient in patients with EOC and lacks biomarkers for selecting the optimal population for immunotherapy (Odunsi, 2017; Lheureux et al., 2019a). Therefore, coping with the challenges and exploiting more effective immunotherapeutic approaches depend on a better understanding of the tumor–immune interactions in the tumor microenvironment (TME) (Mandal and Chan, 2016).

CD38 is a 45-kDa type II transmembrane glycoprotein with ectoenzymatic functions, defined as an ectoenzyme, which participates in the catabolism of nicotinamide adenine dinucleotide (NAD^+) to ADP-ribose and cyclic ADP-ribose (Niels et al., 2018; Hogan et al., 2019), thus playing an important role in adenosinergic pathways and mediating NAD^+ homeostasis. In addition, CD38 has also been described as a surface differentiation marker for lymphocytes, including plasma cells, myeloid cells, and other lymphoid cells (Hogan et al., 2019; Joosse et al., 2019). Because CD38 is uniformly and highly expressed on myeloma cells, a novel therapeutic strategy has emerged that involves targeting CD38 in multiple myeloma; basic research and clinical trials have demonstrated that anti-CD38 mAbs (such as daratumumab) have high efficacy and favorable safety as immunotherapies to increase survival for multiple myeloma patients (Dimopoulos et al., 2016; Horenstein et al., 2019). Recently, studies have also demonstrated that CD38 is involved in $CD8^+$ T-cell suppression via adenosine receptor signaling in the TME, which can cause resistance to PD-1/PD-L1 blockade therapy (Chen et al., 2018). These results showed that CD38 plays multifaceted functional roles in lymphocytes and in the TME. However, the underlying immunological mechanisms and prognostic value of CD38 in the microenvironment of EOC are still unclear.

Here, we used online databases, such as Gene Expression Profiling Interactive Analysis (GEPIA), Oncomine, TISIDB, and Kaplan–Meier plotter (**Supplementary Table S1**), to validate that CD38 was highly expressed in EOC compared with normal ovarian tissue and positively correlated with good prognosis. CD38 was correlated with tumor-infiltrating lymphocytes (TILs), especially with activated $CD8^+$ T cells. These findings uncover the important immunoregulatory role of CD38 in the EOC microenvironment and provide a potential target for ovarian cancer immunotherapy.

MATERIALS AND METHODS

GEPIA Database Analysis

Gene Expression Profiling Interactive Analysis[1] is a comprehensive web-based analysis tool that includes tumor and normal sample RNA sequencing data from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression projects and provides analysis of the interactive relationship, functions, and prognostic value of gene expression in cancer and normal tissues (Tang et al., 2017). The mRNA expression level and prognostic predictive significance of the CD38 gene in EOC were determined in GEPIA. Moreover, gene expression correlation analysis was also conducted by using the GEPIA database.

Oncomine Database Analysis

Oncomine[2] is a gene chip–based online database (Rhodes et al., 2004) that was employed to further verify the expression level of CD38 in EOC.

TISIDB Database Analysis

TISIDB[3] is an integrated repository web portal for analysis of interactions between tumors and the immune system (Ru et al., 2019). It integrates multiple types of data resources in oncoimmunology, including literature mining results from the PubMed database and TCGA. The TISIDB was used to assess the role of CD38 in tumor–immune interplay.

Kaplan–Meier Plotter Database Analysis

Kaplan–Meier plotter[4] is an online database integrating gene expression data and clinical information (Gyorffy et al., 2012). To evaluate the prognostic value of CD38 mRNA expression in ovarian cancer, CD38 was entered into this database to obtain Kaplan–Meier survival plots. The hazard ratio (HR) with 95% confidence intervals and log-rank *P* values were calculated on the web page.

The Tumor Immune Estimation Resource Database Analysis

The Tumor Immune Estimation Resource (TIMER)[5] is a user-friendly web interface for investigating the molecular characterization of tumor–immune interactions (Li et al., 2017). TIMER adopts a deconvolution of previously published computational approaches for estimating the abundance of TILs from gene expression profiles. Approximately six subsets of TILs were pre-calculated in 32 cancer types and data from the TCGA database. The correlations between CD38 mRNA expression and gene markers of TILs were analyzed via correlation modules in TIMER.

TCGA Data Downloading

The level 3 gene expression profile for EOC using Affymetrix HT Human Genome U133a (version September 8, 2017) was downloaded from TCGA datasets[6]. Meanwhile, clinicopathological and survival information were also obtained from the TCGA data portal. The ESTIMATE algorithm (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) was used to calculate immune

[1]http://gepia.cancer-pku.cn/index.html

[2]http://www.oncomine.org

[3]http://cis.hku.hk/TISIDB

[4]www.kmplot.com

[5]https://cistrome.shinyapps.io/timer

[6]https://tcga-data.nci.nih.gov/tcga/

scores and stromal scores of ovarian cancer by applying the downloaded data. The ESTIMATE algorithm was designed by Yoshihara et al. This algorithm can analyze specific gene expression signatures of immune and stromal cells to calculate immune and stromal scores (Yoshihara et al., 2013) and finally predict the non-tumor cell infiltration level.

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was performed to identify significantly enriched groups of genes (Subramanian et al., 2005). In this study, GSEA software[7] was applied to analyze biological pathway divergences between high and low CD38 mRNA in the EOC expression profiles of TCGA data. $P < 0.05$ and FDR (false discovery rate) $q < 0.05$ were considered threshold values to estimate statistical significance.

Calculation of Immune and Stromal Scores

The Cancer Genome Atlas level 3 gene expression data and clinical information were acquired from the Genomic Data Commons (GDC, available at https://portal.gdc.cancer.gov/) data portal on May 10, 2019. Immune and stromal scores were calculated by the ESTIMATE algorithm of the downloaded data for each ovarian cancer sample (Yoshihara et al., 2013). The cutoff values were defined with median scores, and based on the cutoff value, samples were divided into low and high immune/stromal score groups. The survival analysis was assessed by the log-rank test. $P < 0.05$ was considered statistically significant.

Statistical Analysis

Survival analysis of CD38 in EOC was performed by using Kaplan–Meier plotter and GEPIA, and these two databases used the log-rank test for hypothesis evaluation. The Cox proportional hazard ratio and the 95% confidence interval are displayed in the survival curves. The thresholds for high-/low-expression-level cohorts were defined as the median CD38 mRNA level. The correlation of CD38 mRNA expression was assessed by using TIMER and TISIDB. Spearman correlation was calculated, and $P < 0.05$ indicated statistically significant differences.

RESULTS

Expression Levels of CD38 mRNA in EOC

Based on the data of the GEPIA database, the CD38 mRNA levels in EOC and normal ovarian tissues were assessed. The results showed that the CD38 expression level in EOC was higher than that in normal ovarian tissue (**Figure 1A**). In addition, when compared to the different stages of EOC in some data sets, higher expression was observed in stage II, and lower expression was observed in stages III and IV (**Figure 1B**). Unfortunately, data about stage I disease were not found. We further used the Oncomine database to examine CD38 expression in multiple histological types of EOC. This analysis revealed that CD38 mRNA was more highly expressed in malignant EOC than in borderline tumors, and ovarian endometrioid carcinoma had lower CD38 expression than ovarian serous cancer (**Supplementary Figure S1**).

Four molecular subtypes (mesenchymal, immunoreactive, differentiated, and proliferative) have been identified in EOC (Konecny et al., 2014). In TISIDB, we found that CD38 expression was highest in the immunoreactive subtype and lowest in the proliferative subtype (**Figure 1C**). This result implied that CD38 was strongly linked to the tumor immune microenvironment. Shmulevich's study clustered six immune subtypes for cancer (Thorsson et al., 2018). In TISIDB, we further analyzed CD38 expression in different immune subtypes of EOC. We found CD38 was expressed in four types, including C1 (wound healing type), C2 [interferon γ (IFN-γ) dominant type], C3 (inflammatory type), and C4 (lymphocyte depleted type). CD38 was highest in the C2 (IFN-γ dominant) type and lowest in the C3 (inflammatory) type (**Figure 1D**).

The Prognostic Value of CD38 in EOC

The GEPIA database was used to evaluate the correlation of CD38 gene expression with the prognosis of ovarian cancer patients, and this analysis included 424 EOC cases. This analysis revealed that high levels of CD38 (above median) expression were associated with significantly longer disease-free survival (DFS, HR = 0.66, $P = 0.00089$) and increased OS (HR = 0.67, $P = 0.0016$) (**Figures 2A,B**).

To validate CD38 gene expression analysis, we next used the Kaplan–Meier plotter database to investigate the prognostic potential of CD38 expression in EOC, and this analysis included 1,657 patients with OS data and 1,435 patients with progression-free survival (PFS) data. CD38 gene expression was also strongly correlated with increased OS [HR = 0.75 (0.64–0.86), $P = 0.0004$] and PFS [HR = 0.8 (0.73–0.97), $P = 0.0178$] (**Figures 2C,D** and **Table 1**). The detailed relationships between CD38 mRNA expression and prognosis of EOC based on different clinicopathological characteristics in the Kaplan–Meier plotter database are presented in **Table 1**.

In Kaplan–Meier plotter databases, except the microarray analysis of CD38 expression, RNA sequencing data were also acquired and used for online analysis of the prognostic value of CD38 in 373 patients of EOC with diverse tumor mutation statuses. We found that CD38 levels were positively correlated with OS in patients with both high and low mutation burden ($P = 0.0044$ and 0.0027, respectively; **Figures 2E,F**).

The Correlation of CD38 With Immune and Stromal Scores in EOC

The gene expression and clinical data profiles of 469 ovarian serous cystadenocarcinoma patients were downloaded from the TCGA database on May 10, 2019. The ESTIMATE algorithm was applied to assess stromal and immune cells in ovarian cancer. The analysis results implied that stromal scores of EOC were distributed from -1,988.05 to 1,837.43, and immune scores ranged from -1,498.58 to 2,774.16. To determine the potential relevance of CD38 with immune scores and/or stromal

[7] http://www.broadinstitute.org/gsea/

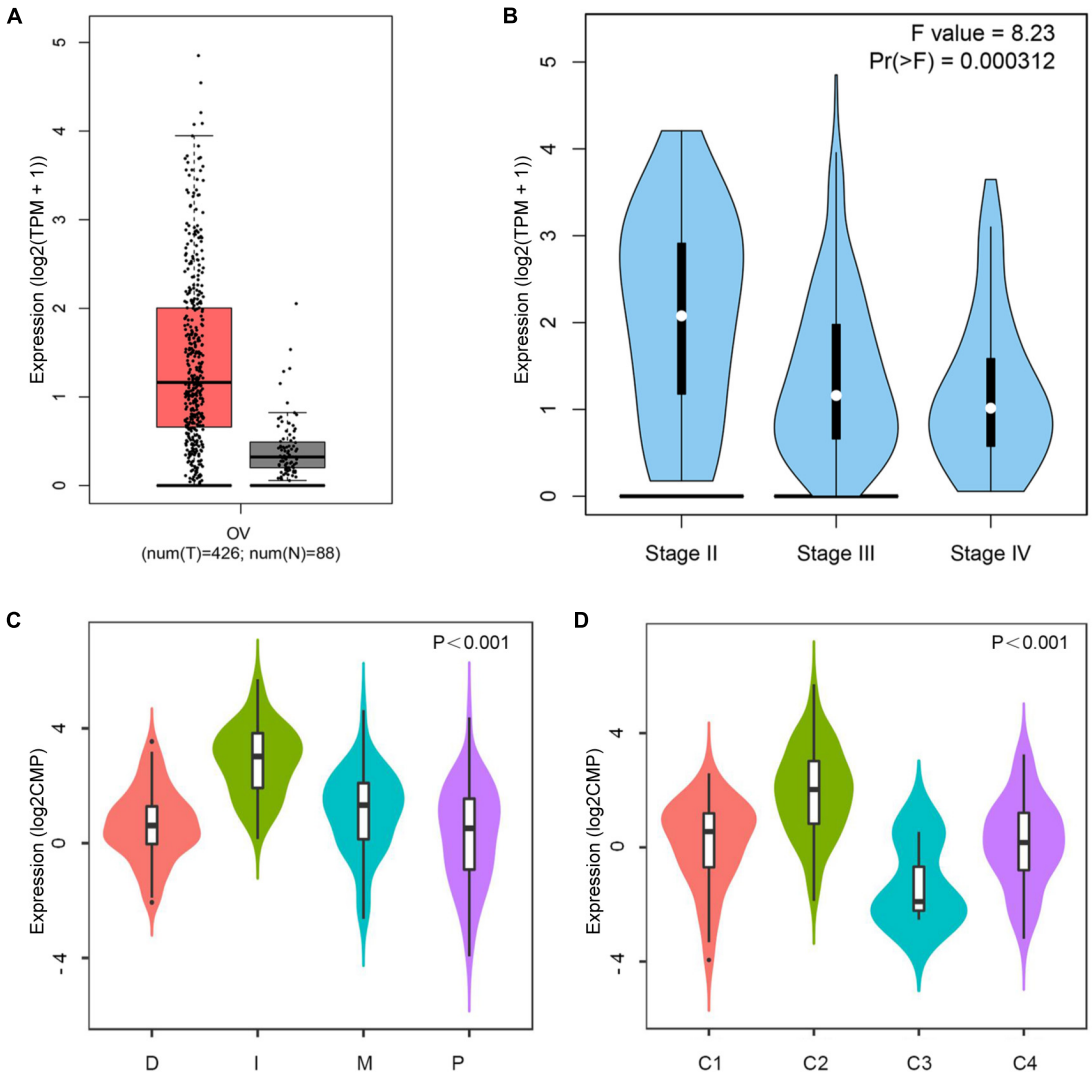


FIGURE 1 | CD38 expression levels in ovarian cancer. **(A)** Epithelial ovarian cancer compared with normal tissues in the GEPIA database. **(B)** CD38 expression levels in different stages of epithelial ovarian cancer from GEPIA database. **(C)** CD38 expression levels in different molecular subtypes of epithelial ovarian cancer from TISIDB database. **(D)** CD38 expression levels in different immune subtypes of epithelial ovarian cancer from TISIDB database. The CD38 gene expression profiles were normalized by log2(TPM + 1) in **(A,B)**, and log counts per million mapped reads (log2CPM) in **(C,D)**.

scores, 469 patients were classified into top (high group) and bottom halves (low group) according to their scores. Patients with high immune scores had higher CD38 expression compared with patients with low immune scores (**Figure 3A**). Consistently, patients with high stromal scores also showed higher CD38 expression compared with patients with low stromal scores (**Figure 3B**).

We further evaluated the prognostic impact of CD38 on the different statuses of immune scores and/or stromal scores for ovarian cancer. For the immune scores, CD38 gene expression was positively correlated with OS of EOC in both the high (above median) immune score group and the low score group (**Figures 3C,D**). The difference was that, for the stromal scores, CD38 gene expression was positively correlated with the OS of EOC in patients with high (above median) stromal scores but not in patients with low stromal scores (**Figures 3E,F**).

CD38 Expression Is Involved in Antitumor Immunity

To further study the roles of CD38 expression in the ovarian cancer microenvironment. Gene set enrichment analysis was conducted by utilizing the gene expression profiles of 469 EOC samples acquired from TCGA database, which contain RNA sequencing data. The gene signatures implied enrichment in many categories, such as immune response, adaptive immune response, lymphocyte activation, regulation of T cell–mediated immunity, and natural killer cell–mediated cytotoxicity, and were enriched in EOC samples with high CD38 expression

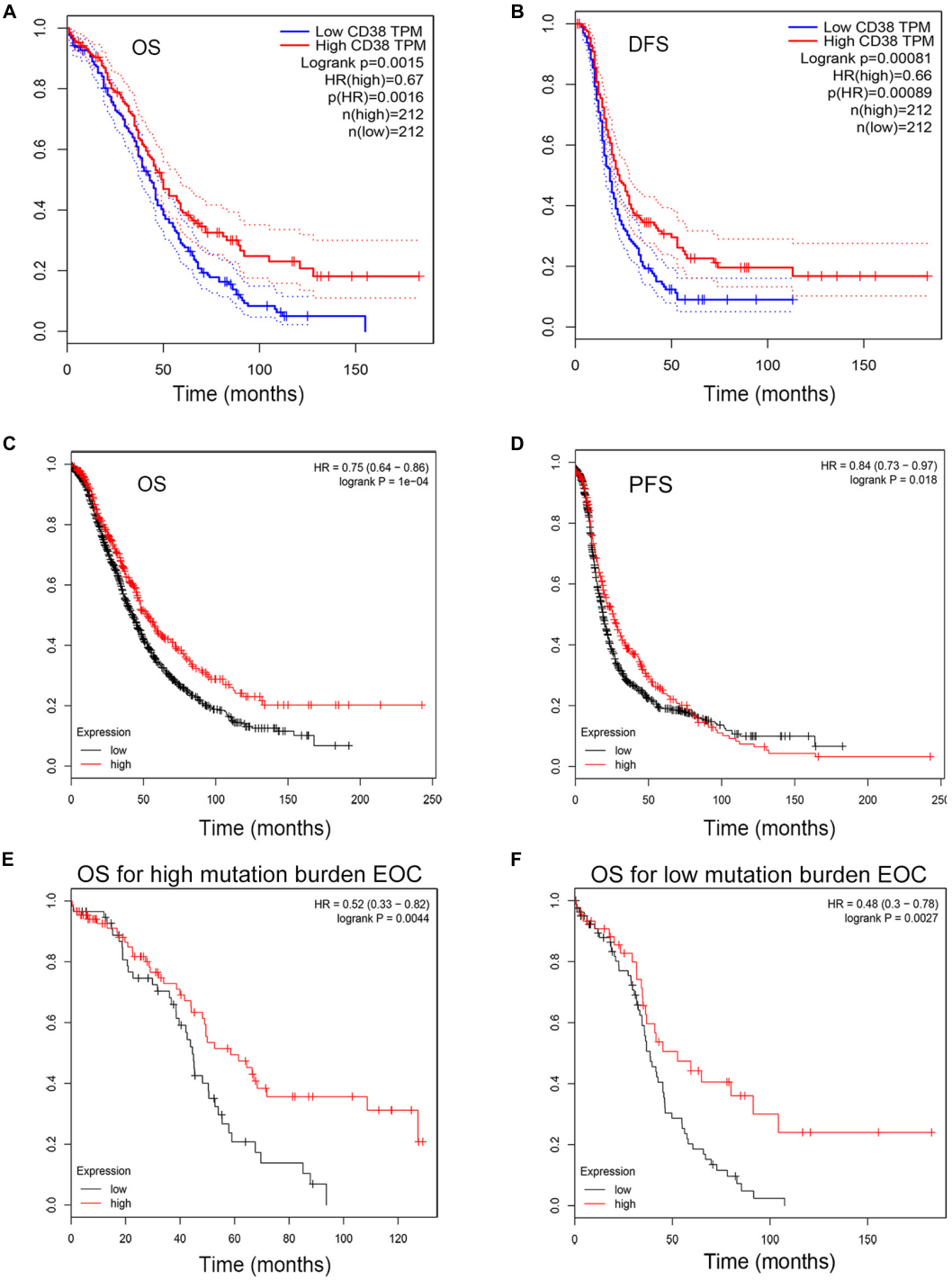


FIGURE 2 | Kaplan–Meier survival curves comparing the high and low expression of CD38 in epithelial ovarian cancer in the GEPIA and Kaplan–Meier plotter databases. **(A,B)** Survival curves of OS and DFS in ovarian cancer from GEPIA databases. **(C,D)** Survival curves of OS and PFS in epithelial ovarian cancer from Kaplan–Meier plotter databases. **(E,F)** High CD38 expression was correlated with better OS either in high or low tumor mutation burden from Kaplan–Meier plotter databases.

TABLE 1 | Correlation of CD38 mRNA expression and clinical prognosis in ovarian cancer with different clinicopathological factors by Kaplan–Meier plotter.

	OS			PFS		
Clinicopathological traits	**n**	**HR**	***P***	**n**	**HR**	***P***
Total	1,656	0.75 (0.64 – −0.86)	1E-04	1,435	0.84 (0.73 – −0.97)	0.0178
Average CA-125 below lower quartile	395	0.6 (0.46 – −0.78)	0.00012	326	0.5 (0.38 – −0.66)	8.5E-07
HISTOLOGY						
Endometrioid	37	5.31 (0.88 – −31.88)	0.041	51	2.53 (1.0 – −6.45)	0.0431
Serous	1,207	0.7 (0.6 – −0.82)	1.3E-05	1,104	0.87 (0.76 – −1.01)	0.0639
STAGE						
I	74	1.74 (0.52 – −5.87)	0.3639	96	5.88 (1.6 – −21.64)	0.0028
II	61	1.91 (0.62 – −5.84)	0.2513	67	1.72 (0.85 – −3.49)	0.1287
III	1,044	0.69 (0.59 – −0.82)	2.2E-5	919	0.83 (0.71 – −0.97)	0.02
IV	176	0.67 (0.45 – −1.0)	0.0488	162	1.37 (0.9 – −2.07)	0.1357
GRADE						
I	56	1.62 (0.61 – −4.31)	0.3328	37	3.29 (1.1 – −9.83)	0.0236
II	324	0.69 (0.49 – −0.97)	0.0295	256	1.3 (0.93 – −1.82)	0.1211
III	1,015	0.65 (0.55 – −0.77)	8.7E-07	837	0.83 (0.7 – −0.99)	0.0346
P53						
Mutated	506	0.7 (0.55 – −0.89)	0.0043	483	0.71 (0.56 – −0.89)	0.0025
Wild type	94	1.36 (0.74 – −2.48)	0.318	84	1.55 (0.88 – −2.72)	0.1223
DEBULK						
Optimal	801	0.67 (0.52 – −0.87)	0.0022	696	0.84 (0.7 – −1.02)	0.0719
Suboptimal	459	0.72 (0.59 – −0.88)	0.0011	459	0.67 (0.54 – −0.83)	0.0002
CHEMOTHERAPY						
Contains platin	1,409	0.75 (0.65 – −0.86)	4.8E-05	1,259	0.75 (0.66 – −0.86)	2.2E-05
Contains Taxol	793	0.62 (0.49 – −0.78)	5.8E-05	715	0.79 (0.65 – −0.95)	0.0126
Contains Avastin	50	0.55 (0.21 – −1.43)	0.2168	50	0.75 (0.39 – −1.45)	0.391

Bold values indicate P < 0.05.

(**Figure 4**). This analysis revealed that CD38 might play vital roles in antitumor immune modulation.

The Relationship Between CD38 Expression and Immune Cell Infiltration

Several studies have implied that TILs are a prognostic indicator for ovarian cancer (Zhang et al., 2003). Therefore, the associations between CD38 gene expression and TILs infiltration level in EOC were analyzed in the TIMER database. This analysis showed that CD38 was significantly correlated with tumor purity, $CD8^+$ T cells, $CD4^+$ T cells, and B cells in EOC. Myeloid cell types, including macrophages, neutrophils, and dendritic cells, were also significantly correlated with CD38 expression (**Figure 5A**). In the TISIDB database, we also found that CD38 was strongly related to immune infiltration in EOC, especially the infiltration of activated immune cells, such as activated $CD8^+$ T cells (R = 0.68), activated $CD4^+$ T cells (R = 0.604), and activated B cells (R = 0.663) (**Figures 5B–D** and **Supplementary Table S2**). Interestingly, the relationship between CD38 and memory immune cells was not strong (**Figure 5E** and **Supplementary Table S3**). To further clarify the relationship between CD38 and various subtypes of TILs in ovarian cancer, the TIMER and TISIDB online databases were employed to further analyze the relationship between CD38 and marker genes of different immune cells, including $CD8^+$ T cells, $CD4^+$ T cells, B cells, macrophages, neutrophils, and dendritic cells in EOC (**Table 2** and **Supplementary Table S3**).

DISCUSSION

As a multifunctional ADP-ribosyl cyclase, CD38 is widely expressed on plasma cells and other types of immune cells (Deaglio et al., 2001). With daratumumab (an anti-CD38 mAb) approved for clinical application, CD38 has emerged as a high-impact therapeutic target in multiple myeloma (Nijhof et al., 2015; Elsada and Adler, 2019). The CD38/CD203a/CD73 adenosinergic pathway is a major regulatory mechanism in niche metabolic reprogramming (Horenstein et al., 2013). Furthermore, CD38 is expressed on various lymphocytes, including regulatory T cells (Tregs), B cells, and myeloid cells, which have potential immunomodulatory effects (Flores-Borja et al., 2013; Karakasheva et al., 2015; Feng et al., 2017). However, the role of immunologic reprogramming in the solid TME is still unclear. Here, we present a study that revealed that CD38 expression levels correlate with prognosis in ovarian cancer. High expression of CD38 correlates with early disease stage and better prognosis. In addition, our analyses show that TILs and diverse immune markers in ovarian cancer are associated with CD38 expression levels. Hence, our comprehensive and systematic analysis study provides

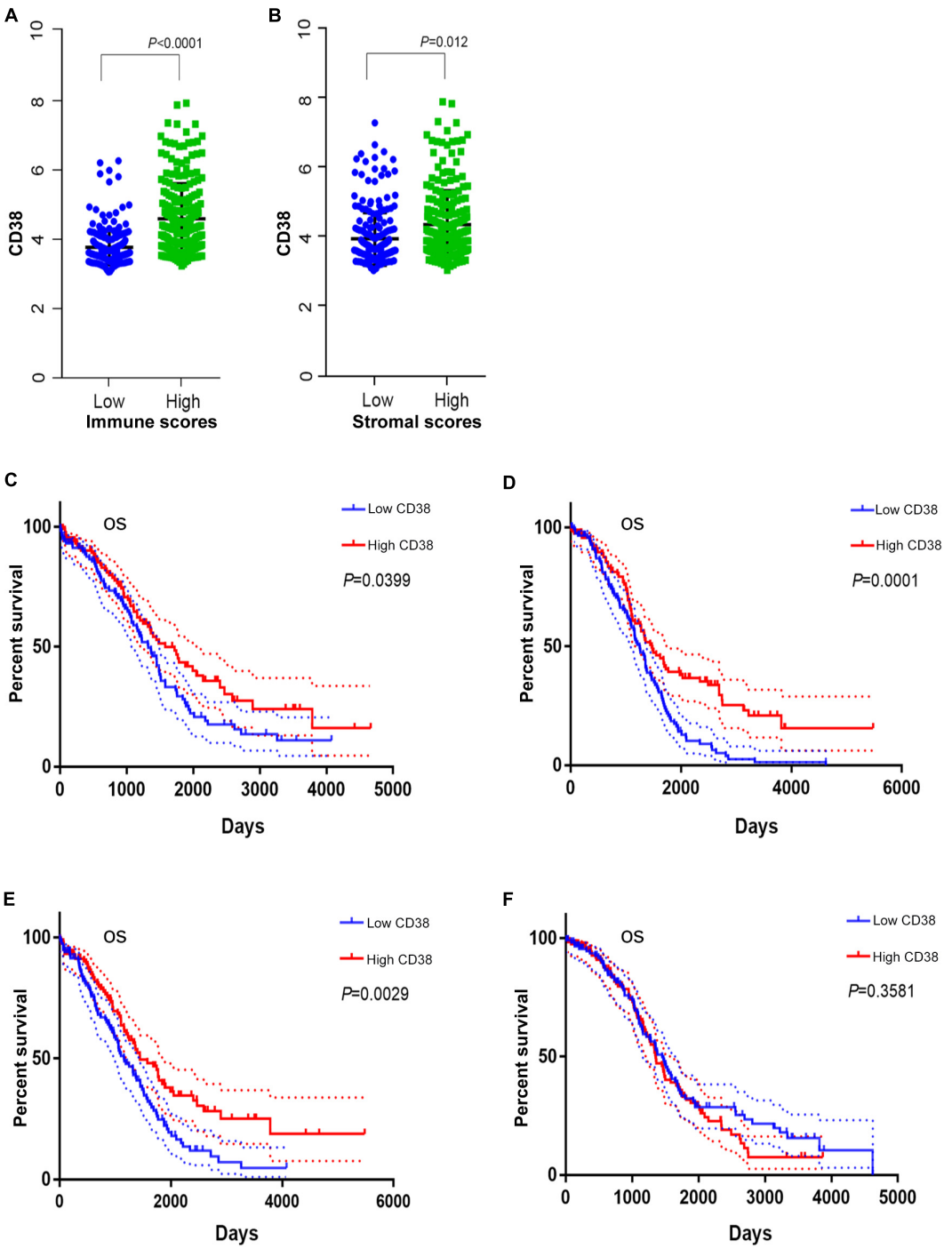


FIGURE 3 | The correction between CD38 and immune or stromal scores (this analysis in ovarian cancer patients with immune or stromal scores median cutoff). **(A)** CD38 highly expressed in high immune scores group from TCGA database. **(B)** CD38 highly expressed in high stromal scores group from TCGA database. **(C)** Survival curves of OS in high immune scores group of epithelial ovarian cancer from TCGA database. **(D)** Survival curves of OS in low immune scores group of epithelial ovarian cancer from TCGA database. **(E)** Survival curves of OS in high stromal scores group of epithelial ovarian cancer from TCGA database. **(F)** Survival curves of OS in low stromal scores group of epithelial ovarian cancer from TCGA database.

valuable insights into the potential immune regulatory role of CD38 in the EOC niche and suggests its use as a cancer prognostic biomarker.

Our study analyzed the CD38 mRNA expression level in normal ovaries and EOC by using online datasets in GEPIA, Oncomine, and TISIDB. The expression of the CD38 gene

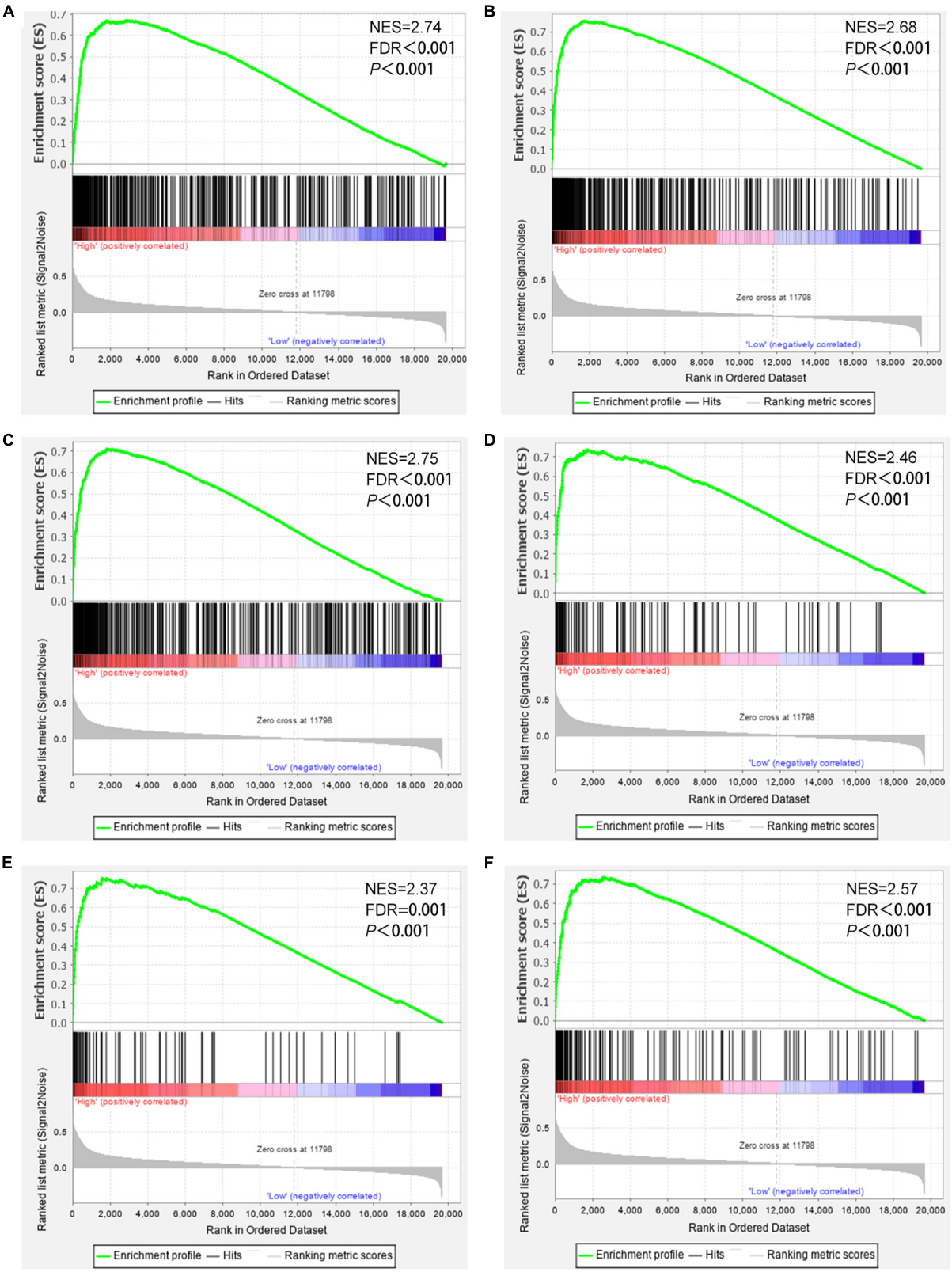


FIGURE 4 | Gene set enrichment analysis showed that CD38 expression is involved in ovarian cancer patients' antitumor immune responses. **(A)** Gene sets representing Innate immune response. **(B)** Adaptive immune response. **(C)** Lymphocyte activation. **(D)** Positive regulation of lymphocyte mediated immunity. **(E)** Regulation of T cell–mediated immunity. **(F)** Natural killer cell–mediated cytotoxicity.

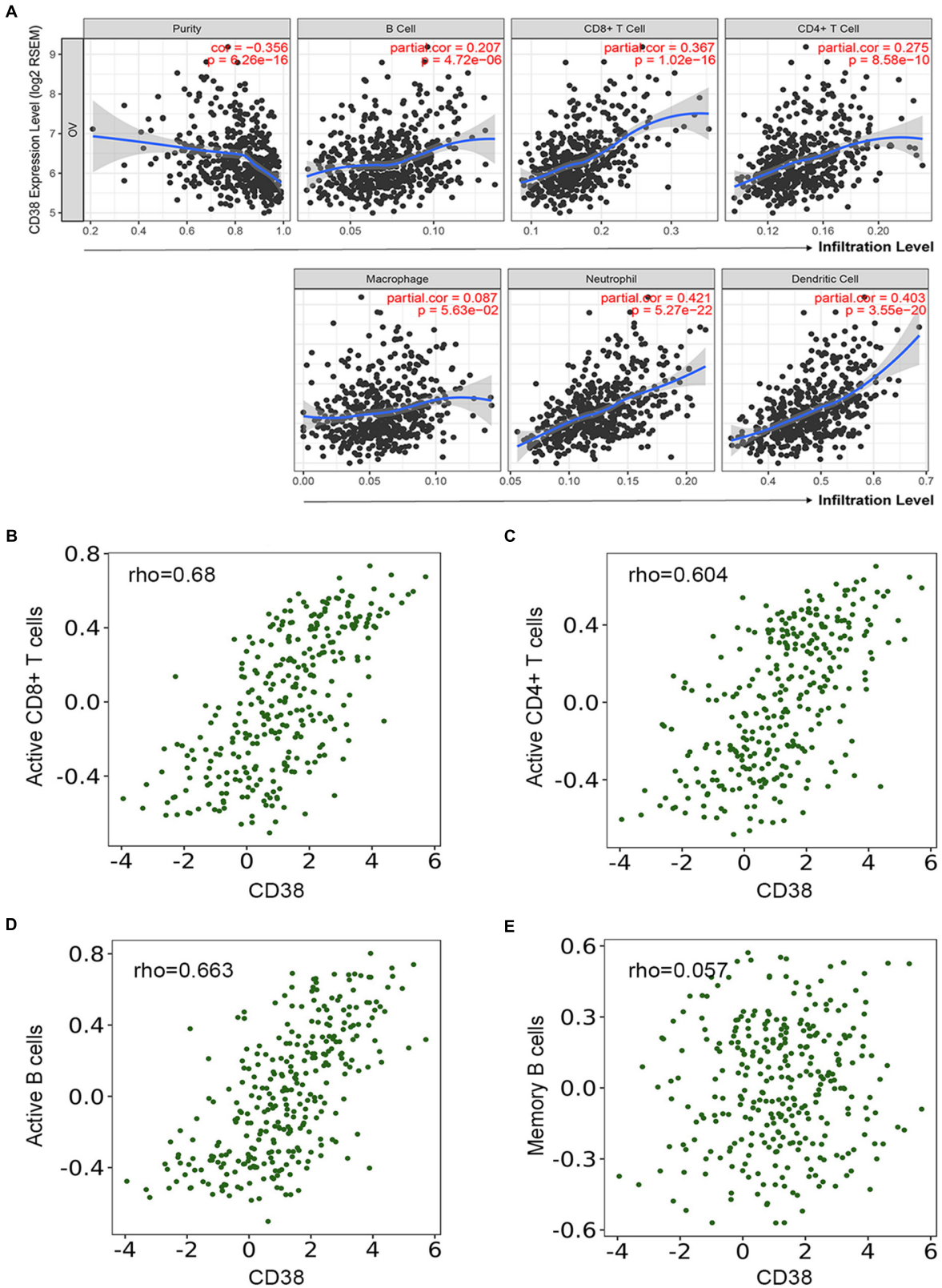


FIGURE 5 | Correlation of CD38 expression with immune infiltration level in epithelial ovarian cancer. **(A)** CD38 expression is significantly negatively related to tumor purity and has significant positive correlations with infiltrating levels of B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells from TIMER database. **(B–E)** CD38 expression has significant positive correlations with active $CD8^+$ T cells, active $CD4^+$ T cells, and active B cells, other than memory B cells.

TABLE 2 | Correlation analysis between CD38 and relate genes and markers of immune cells in TIMER.

		None		Purity	
Immune profile	**Immune gene**	**Cor**	***P***	**Cor**	***P***
T CELL					
	CD3D	0.695	5.34E-45	0.654	1.03E-31
	CD3E	0.711	0.00	0.687319	3.71E-36
	CD3G	0.627	1.89E-34	0.543433	1.56E-20
	CD2	0.737	4.36E-53	0.718886	6.87E-41
$CD4^+$ T CELL					
	CD4	0.535	0.00	0.468871	5.16E-15
$CD8^+$ T CELL					
	CD8A	0.656	0.00	0.584927	2.98E-24
	CD8B	0.539	0.00	0.448	1.01E-13
	TBX21	0.738	2.51E-53	0.727225	2.99E-42
	EOMES	0.586	2.14E-29	0.512907	4.14E-18
	LCK	0.604	0.00	0.559081	7.10E-22
	IFNG	0.674	1.88E-41	0.610228	8.56E-27
	PRF1	0.687	0.00	0.663527	5.71E-33
	GZMA	0.655	0.00	0.629182	7.47E-29
	GZMB	0.665	5.01E-40	0.633114	2.68E-29
	GZMH	0.643	9.21E-37	0.59184	6.34E-25
	GZMK	0.625	3.56E-34	0.562898	3.26E-22
	GZMM	0.647	2.45E-37	0.615405	2.42E-27
	CXCL9	0.704	0.00	0.642913	1.95E-30
	CXCL10	0.792	0.00	0.767956	1.07E-49
T_H1					
	IFNG	0.674	1.88E-41	0.610228	8.56E-27
	TBX21	0.738	2.51E-53	0.727225	2.99E-42
	TNF	0.247	1.46E-05	0.176	5.45E-03
	STAT4	0.673	2.98E-41	0.624	2.78E-28
	STAT1	0.641	0.00	0.61	1.02E-26
T_H2					
	GATA3	0.277	1.04E-06	0.114	7.32E-02
	STAT6	0.066	2.55E-01	0.061	3.37E-01
	STAT5A	0.224	8.82E-05	0.216	5.84E-04
	IL13	0.167	3.48E-03	0.143	2.42E-02
Tfh					
	CXCR5	0.424	9.16E-20	0.338595	4.28E-08
	CXCL13	0.693	1.31E-44	0.618193	1.21E-27
	BCL6	−0.004	9.43E-01	0.067	2.93E-01
	IL21	0.325	7.11E-09	0.314	4.16E-07
T_H17					
	IL17A	0.151	8.27E-03	0.118106	0.062768
	RORC	−0.114	4.82E-02	−0.04595	0.470446
	IL23A	0.076	1.88E-01	0.085397	0.179202
	STAT3	0.232	4.66E-05	0.153	1.56E-02
Treg					
	FOXP3	0.663	0.00	0.604827	3.12E-26
	IKZF2	−0.092	1.09E-01	−0.07343	0.248332
	IL10	0.321	1.31E-08	0.206341	0.001057
	TGFB1	0.362	1.09E-10	0.194492	0.002049
	CCR8	0.49	9.65E-20	0.408	2.03E-11
	STAT5B	−0.061	2.90E-01	−0.079	2.13E-01
CHECKPOINTS					
	CTLA4	0.738	2.14E-53	0.70801	3.46E-39
	PDCD1	0.609	3.67E-32	0.558032	8.78E-22
	LAG3	0.764	0.00	0.750288	2.73E-46
	PDL1/CD274	0.682	0.00	0.642702	2.07E-30
	TIM3/HAVCR2	0.578	0.00	0.512411	4.51E-18
	TIGIT	0.733	3.26E-52	0.694518	3.50E-37
PROINFLAMMATION					
	PTGS2	0.11	5.67E-02	−0.00507	0.936513
	IL8	0.081	1.59E-01	0.006259	0.921719
	IL1A	0.098	8.77E-02	0.02809	0.659131
	IL1B	0.305	7.07E-08	0.169372	0.007393
	IL18	0.32	1.48E-08	0.245112	9.30E-05
	IL6	0.273	1.36E-06	0.134232	0.034253
	IL12A	0.223	9.37E-05	0.179456	0.004503
	TNF	0.247	1.46E-05	0.175629	0.005451
METABOLISM					
	IDO1	0.584	0.00	0.489875	1.96E-16
	NOS2	−0.021	7.13E-01	−0.08555	0.178423
	HIF1A	0.01	8.63E-01	−0.07241	0.254943
APC/DC					
	HLA-DPA1	0.559	0.00	0.484698	4.48E-16
	HLA-DPB1	0.519	0.00	0.433577	7.79E-13
	HLA-DQA1	0.474	0.00	0.376692	8.15E-10
	HLA-DRA	0.508	0.00	0.431049	1.09E-12
	HLA-DMA	0.456	0.00	0.392215	1.39E-10
	HLA-DQB1	0.36	1.36E-10	0.275	1.05E-05
	BDCA-1/CD1C	0.191	8.09E04	0.073	2.53E-01
	BDCA-4/NRP1	0.176	2.18E-03	0.038	5.54E-01
	CD11C/ITGAX	0.489	0.00	0.422	3.41E-12
B CELL					
	BLK	0.325	7.28E-09	0.245509	9.05E-05
	CD19	0.352	2.85E-10	0.3409	3.42E-08
	MS4A1	0.57	01.75E-27	0.478648	1.16E-15
	CD79A	0.62	1.43E-33	0.522	8.55E-19
MONOCYTE					
	CD86	0.639	0.00	0.579	1.02E-23
	CD115/CSF1R	0.411	7.85E-14	0.306	8.61E-07
TAM					
	CCL2	0.428	0.00	0.359	5.38E-09
	CD68	0.584	0.00	0.532	1.33E-19
	CSF2	0.338	1.56E-09	0.318	3.03E-07
M1					
	INOS/NOS2	−0.021	7.13E-01	−0.086	1.78E-01
	IRF5	0.263	3.65E-06	0.236	1.75E-04
	COX2/PTGS2	0.11	5.67E-02	−0.005	9.37E-01
M2					
	CD163	0.511	0.00	0.424	2.70E-12
	VSIG4	0.438	0.00	0.33	9.62E-08
	MS4A4A	0.539	0.00	0.484	4.73E-16

(Continued)

TABLE 2 | Continued

Immune profile	Immune gene	None		Purity	
		Cor	*P*	Cor	*P*
N					
	CD66B/CEACAM8	−0.094	1.01E-01	−0.083	1.93E-01
	CD11B/ITGAM	0.454	0.00	0.373	1.25E-09
	CCR7	0.65	0.00	0.614	3.20E-27
NK					
	KIR2DL1	0.225	7.82E-05	0.135	3.37E-02
	KIR2DL3	0.24	2.46E-05	0.212	7.38E-04
	KIR2DL4	0.53	2.67E-23	0.497	6.38E-17
	KIR3DL1	0.392	1.44E-12	0.353	1.03E-08
	KIR3DL2	0.188	9.88E-04	0.134	3.40E-02
	KIR3DL3	0.148	9.94E-03	0.12	5.96E-02
	KIR2DS4	0.19	9.16E-04	0.128	4.32E-02

*Cor, R value of Spearman correlation; None, correlation without adjustment; Purity, correlation adjusted by purity. *P < 0.01; **P < 0.001; ***P < 0.0001.*

in EOC was not only higher than that in normal tissue but was also higher than that in borderline ovarian tumors. Nevertheless, ovarian cancer is not a single disease and can be subdivided into many molecular subtypes. Analysis of the TISIDB database showed that the CD38 gene had the highest expression level in the immunoreactive subtype, followed by the mesenchymal type, with little expression in the differentiated and proliferative types. Different levels of CD38 expression in distinct immune subtypes of ovarian cancer were observed, and the C2 (IFN-γ dominant) type had the highest level compared with the other three subtypes. The comprehensive and detailed analysis of CD38 gene expression in various databases among EOC and different subtypes may reflect that CD38 is strongly linked to immunological properties in the microenvironment.

Nevertheless, in the Kaplan–Meier plotter and GEPIA databases, the analysis found matching prognostic value correlations between CD38 expressions in EOC. The increased CD38 expression correlated with better survival in EOC and was not influenced by the immune scores. In addition, high CD38 expression was related to favorable prognosis of EOC in stages III and IV and grades II and III. Together, these results robustly indicated that CD38 is a potential prognostic biomarker for ovarian cancer.

Another important finding is that CD38 expression is closely related to the immune response and lymphocyte infiltration in EOC. Under physiological conditions, CD38 induced mature B-cell proliferation and immunoglobulin M (IgM) secretion. And in CD38 expressed higher on activated T cells, the $CD38^+$ T cells inhibited $CD38^-$ T-cell proliferation to maintain T-cell homeostasis (Bahri et al., 2012; Glaria and Valledor, 2020). On the contrary, another study have unveiled that T cells expressing high levels of CD38 have an extremely low proliferative ability but an enhanced capacity to produce interleukin 2 (IL-2) and IFN-γ (Sandoval-Montes and Santos-Argumedo, 2005). These evidences all suggested that CD38 plays a vital role in the regulation of immune cells activation and differentiation. But its exact regulatory function still needs further study. The GSEA and correlation analyses in our study implied that CD38 regulated the tumor immune microenvironment in EOC and was associated with B- and T-cell activation and regulated immune responses. A study also certified that in human lung cancer CD38 protein is highly expressed in $CD8^+$ tissue-resident memory cells, $CD103^+$ (T_{RM} cells), and a high density of T_{RM} cell infiltration predicts a better prognosis (Ganesan et al., 2017).

Another study revealed that CD38 is one of the essential mechanisms by which tumors obtain resistance to immune checkpoint blockade immunotherapy, resulting in $CD8^+$ T-cell dysfunction. Interferon β might be a factor increasing CD38 expression in the TME (Chen et al., 2018). In addition, Schietinger et al. certified that $PD1^{hi}$ TILs were a heterogeneous population and that $PD1^{hi}$ T cells with increased CD38 expression did not respond to PD-1 and/or PD-L1 immune checkpoint blockers. $CD38^+$ $PD1^{hi}$ T cells may be in a fixed dysfunctional state rather than the plastic reprogrammable state (Philip et al., 2017). All of the studies hinted that CD38 plays a vital role in remodeling the immune microenvironment, and CD38 deserves further research as an immunotherapeutic target and prognostic biomarker in ovarian cancer.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the GEPIA (http://gepia.cancer-pku.cn/index.html), Oncomine (http://www.oncomine.org), TISIDB (http://cis.hku.hk/TISIDB), Tumor Immune Estimation Resource (TIMER, https://cistrome.shinyapps.io/timer), TCGA databases (https://tcga-data.nci.nih.gov/tcga/).

AUTHOR CONTRIBUTIONS

JZ and ZZ: study concept and design. YZ, ZJ, and YL: acquisition and analysis of the data. JZ, YZ, and ZZ: drafting and revising of the manuscript.

FUNDING

This study was partially supported by the National Natural Science Foundation of China (81902626).

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from Prof. Jiangwen Zhang from TISIDB, and our research team for help during the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00369/full#supplementary-material

FIGURE S1 | CD38 expression levels in different types of epithelial ovarian tumor. **(A)** CD38 in data sets of epithelial ovarian cancer compared with borderline ovarian tumor in the Oncomine database. **(B)** CD38 in data sets of ovarian serous cancer compared with ovarian endometrioid cancer in the Oncomine database.

TABLE S1 | Detailed information of the online databases applied in the study.

TABLE S2 | Spearman correlation analysis between expression of CD38 and TILs in epithelial ovarian cancer from TISIDB database.

TABLE S3 | Spearman correlation analysis between expression of CD38 and Immunomodulator in epithelial ovarian cancer from TISIDB database.

REFERENCES

Bahri, R., Bollinger, A., Bollinger, T., Orinska, Z., and Bulfone-Paus, S. (2012). Ectonucleotidase CD38 demarcates regulatory, memory-like CD8+ T cells with IFN-gamma-mediated suppressor activities. *PLoS One* 7:e45234. doi: 10.1371/journal.pone.0045234

Bowtell, D. D., Bohm, S., Ahmed, A. A., Aspuria, P. J., Bast, R. C. Jr., Beral, V., et al. (2015). Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nature reviews. *Cancer* 15, 668–679. doi: 10.1038/nrc4019

Chen, L., Diao, L., Yang, Y., Yi, X., Rodriguez, B. L., Li, Y., et al. (2018). CD38-Mediated Immunosuppression as a Mechanism of Tumor Cell Escape from PD-1/PD-L1 Blockade. *Cancer Discov.* 8, 1156–1175. doi: 10.1158/2159-8290.CD-17-1033

Deaglio, S., Mehta, K., and Malavasi, F. (2001). Human CD38: a (r)evolutionary story of enzymes and receptors. *Leuk. Res.* 25, 1–12. doi: 10.1016/s0145-2126(00)00093-x

Dimopoulos, M. A., Oriol, A., Nahi, H., San-Miguel, J., Bahlis, N. J., Usmani, S. Z., et al. (2016). Daratumumab, lenalidomide, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* 375, 1319–1331.

Elsada, A., and Adler, A. I. (2019). NICE guidance on daratumumab with bortezomib and dexamethasone for previously treated multiple myeloma. *Lancet Oncol.* 20, 619–620. doi: 10.1016/s1470-2045(19)30222-0

Feng, X., Zhang, L., Acharya, C., An, G., Wen, K., Qiu, L., et al. (2017). Targeting CD38 suppresses induction and function of T regulatory cells to mitigate immunosuppression in multiple myeloma. *Clin. Cancer Res.* 23, 4290–4300. doi: 10.1158/1078-0432.CCR-16-3192

Flores-Borja, F., Bosma, A., Ng, D., Reddy, V., Ehrenstein, M. R., Isenberg, D. A., et al. (2013). CD19+CD24hiCD38hi B cells maintain regulatory T cells while limiting TH1 and TH17 differentiation. *Sci. Transl. Med.* 5:173ra23. doi: 10.1126/scitranslmed.3005407

Ganesan, A. P., Clarke, J., Wood, O., Garrido-Martin, E. M., Chee, S. J., Mellows, T., et al. (2017). Tissue-resident memory features are linked to the magnitude of cytotoxic T cell responses in human lung cancer. *Nat. Immunol.* 18, 940–950. doi: 10.1038/ni.3775

Glaria, E., and Valledor, A. F. (2020). Roles of CD38 in the immune response to infection. *Cells* 9:E228. doi: 10.3390/cells9010228

Gyorffy, B., Lanczky, A., and Szallasi, Z. (2012). Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer* 19, 197–208. doi: 10.1530/ERC-11-0329

Hogan, K. A., Chini, C. C. S., and Chini, E. N. (2019). The Multi-faceted Ecto-enzyme CD38: roles in immunomodulation, cancer, aging, and metabolic diseases. *Front. Immunol.* 10:1187. doi: 10.3389/fimmu.2019.01187

Horenstein, A. L., Bracci, C., Morandi, F., and Malavasi, F. (2019). CD38 in adenosinergic pathways and metabolic re-programming in human multiple myeloma cells: in-tandem insights from basic science to therapy. *Front. Immunol.* 10:760. doi: 10.3389/fimmu.2019.00760

Horenstein, A. L., Chillemi, A., Zaccarello, G., Bruzzone, S., Quarona, V., Zito, A., et al. (2013). A CD38/CD203a/CD73 ectoenzymatic pathway independent of CD39 drives a novel adenosinergic loop in human T lymphocytes. *Oncoimmunology* 2:e26246. doi: 10.4161/onci.26246

Joosse, M. E., Menckeberg, C. L., de Ruiter, L. F., Raatgeep, H. R. C., van Berkel, L. A., Simons-Oosterhuis, Y., et al. (2019). Frequencies of circulating regulatory TIGIT(+)CD38(+) effector T cells correlate with the course of inflammatory bowel disease. *Mucosal Immunol.* 12, 154–163. doi: 10.1038/s41385-018-0078-4

Karakasheva, T. A., Waldron, T. J., Eruslanov, E., Kim, S. B., Lee, J. S., O'Brien, S., et al. (2015). CD38-expressing myeloid-derived suppressor cells promote tumor growth in a murine model of esophageal cancer. *Cancer Res.* 75, 4074–4085. doi: 10.1158/0008-5472.CAN-14-3639

Konecny, G. E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K. R., Dering, J., et al. (2014). Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* 106:dju249. doi: 10.1093/jnci/dju249

Lheureux, S., Braunstein, M., and Oza, A. M. (2019a). Epithelial ovarian cancer: evolution of management in the era of precision medicine. *CA* 69, 280–304. doi: 10.3322/caac.21559

Lheureux, S., Gourley, C., Vergote, I., and Oza, A. M. (2019b). Epithelial ovarian cancer. *Lancet* 393, 1240–1253.

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77, e108–e110. doi: 10.1158/0008-5472.CAN-17-0307

Mandal, R., and Chan, T. A. (2016). Personalized oncology meets immunology: the path toward precision immunotherapy. *Cancer Discov.* 6, 703–713. doi: 10.1158/2159-8290.CD-16-0146

Menon, U., Karpinskyj, C., and Gentry-Maharaj, A. (2018). Ovarian cancer prevention and screening. *Obstet. Gynecol.* 131, 909–927. doi: 10.1097/AOG.0000000000002580

Niels, W. C. J., van de Donk, P. G. R., and Malavasi, F. (2018). CD38 antibodies in multiple myeloma: back to the future. *Blood* 131, 13–29. doi: 10.1182/blood-2017-06-740944

Nijhof, I. S., Groen, R. W., Noort, W. A., van Kessel, B., de Jong-Korlaar, R., Bakker, J., et al. (2015). / Preclinical evidence for the therapeutic potential of CD38-Targeted immuno-chemotherapy in multiple myeloma patients refractory to Lenalidomide and Bortezomib. *Clin. Cancer Res.* 21, 2802–2810. doi: 10.1158/1078-0432.CCR-14-1813

Odunsi, K. (2017). Immunotherapy in ovarian cancer. *Ann. Oncol.* 28, 81–87.

Philip, M., Fairchild, L., Sun, L., Horste, E. L., Camara, S., Shakiba, M., et al. (2017). Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* 545, 452–456. doi: 10.1038/nature22367

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6, 1–6. doi: 10.1016/s1476-5586(04)80047-2

Ribas, A., and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science* 359, 1350–1355. doi: 10.1126/science.aar4060

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* 35, 4200–4202. doi: 10.1093/bioinformatics/btz210

Sandoval-Montes, C., and Santos-Argumedo, L. (2005). CD38 is expressed selectively during the activation of a subset of mature T cells with reduced proliferation but improved potential to produce cytokines. *J. Leukoc. Biol.* 77, 513–521. doi: 10.1189/jlb.0404262

Scott, G. L. H., Jackman, D., Spigel, D., Antonia, S., Hellmann, M., Powderly, J., et al. (2018). Five-Year Follow-Up of Nivolumab in Previously treated advanced non-small-cell lung cancer: results from the CA209-003 study. *J. Clin. Oncol.* 36, 1675–1684. doi: 10.1200/JCO.2017.77.0412

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2018). The immune landscape of cancer. *Immunity* 48:e14.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). / Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

Zhang, L., Conejo-Garcia, J. R., Katsaros, D., Gimotty, P. A., Massobrio, M., Regnani, G., et al. (2003). Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N. Engl. J. Med.* 348, 203–213. doi: 10.1056/nejmoa020177

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer XC and handling Editor declared their shared affiliation.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 May 2020
doi: 10.3389/fgene.2020.00420

OSluca: An Interactive Web Server to Evaluate Prognostic Biomarkers for Lung Cancer

*Zhongyi Yan[1†], Qiang Wang[1†], Zhendong Lu[1†], Xiaoxiao Sun[1], Pengfei Song[1], Yifang Dang[1], Longxiang Xie[1], Lu Zhang[1], Yongqiang Li[1], Wan Zhu[2], Tiantian Xie[3], Jing Ma[3], Yijie Zhang[3] and Xiangqian Guo[1]**

[1] Department of Predictive Medicine, Institute of Biomedical Informatics, Cell Signal Transduction Laboratory, Bioinformatics Center, Henan Provincial Engineering Center for Tumor Molecular Medicine, School of Software, School of Basic Medical Sciences, Henan University, Kaifeng, China, [2] Department of Anesthesia, Stanford University, Stanford, CA, United States, [3] Department of Respiratory and Critical Care Medicine, Huaihe Hospital of Henan University, Kaifeng, China

Lung cancer is the principal cause of leading cancer-related incidence and mortality in the world. Various studies have excavated the potential prognostic biomarkers for cancer patients based on gene expression profiles. However, most of these reported biomarkers lack independent validation in multiple cohorts. Herein, we collected 35 datasets with long-term follow-up clinical information from TCGA (2 cohorts), GEO (32 cohorts), and Roepman study (1 cohort), and developed a web server named OSluca (Online consensus Survival for Lung Cancer) to assess the prognostic value of genes in lung cancer. The input of OSluca is an official gene symbol, and the output web page of OSluca displays the survival analysis summary with a forest plot and a survival table from Cox proportional regression in each cohort and combined cohorts. To test the performance of OSluca, 104 previously reported prognostic biomarkers in lung carcinoma were evaluated in OSluca. In conclusion, OSluca is a highly valuable and interactive prognostic web server for lung cancer. It can be accessed at http://bioinfo.henu.edu.cn/LUCA/LUCAList.jsp.

Keywords: survival, lung cancer, biomarker, prognosis, OSluca

OPEN ACCESS

Edited by:
Harinder Singh,
J. Craig Venter Institute, United States

Reviewed by:
Sipeng Shen,
Nanjing Medical University, China
Sudipto Saha,
Bose Institute, India

****Correspondence:***
Xiangqian Guo
xqguo@henu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *30 October 2019*
Accepted: *03 April 2020*
Published: *26 May 2020*

Citation:
Yan Z, Wang Q, Lu Z, Sun X, Song P, Dang Y, Xie L, Zhang L, Li Y, Zhu W, Xie T, Ma J, Zhang Y and Guo X (2020) OSluca: An Interactive Web Server to Evaluate Prognostic Biomarkers for Lung Cancer. Front. Genet. 11:420. doi: 10.3389/fgene.2020.00420

INTRODUCTION

Lung cancer (LUCA) is an aggressive disease with leading mortality and incidence in the world. Based on histology, there are two types of LUCA, including non-small cell lung cancer (NSCLC), which accounts for 80% of LUCA and small cell lung cancer (SCLC), which accounts for approximately 20% of LUCA (Raponi et al., 2006; Bray et al., 2018). NSCLC can be further sub-divided into four subtypes, including adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and bronchioloalveolar carcinoma (Ramalingam et al., 2011). Classical histological subtypes indeed play a dominant role in treatment and prognosis of lung cancer. Recently, reclassification of lung cancer based on tumor biomarkers improves lung cancer therapy (Beer et al., 2002; Hoadley et al., 2018).

Many studies have demonstrated that using clinical-association-prognostic biomarkers can assist the characterization of cancer subtypes and provide new insights of cancer recurrence and patients response to more precise therapies (Meyerson and Carbone, 2005; Bild et al., 2006;

Raponi et al., 2006). It is worth noting that numerous single- or multi-prognostic biomarkers have been identified using high-throughput profiling methods (Raponi et al., 2006). By mining a mass of these profiling data deposited in public database, meta-analysis has exploited potential prognostic genes, such as *KRT8* (Xie et al., 2019a). However, for biologists and clinicians, it is technically difficult to analyze these massive public data to screen and develop prognostic biomarkers. Previously, we have built several web servers of prognostic biomarker analysis for breast cancer, esophageal carcinoma, etc. (Wang et al., 2019a,b,c, 2020; Xie et al., 2019b,c; Yan et al., 2019; Zhang et al., 2019, 2020; Dong et al., 2020). In this current study, we have integrated bulky RNA expression profiles of lung cancer with clinical survival information, mainly from TCGA (The Cancer Genome Atlas) and GEO (the Gene Expression Omnibus) databases, and built a prognostic analysis web server named OSluca (Online consensus Survival for Lung Cancer) to analyze and evaluate prognostic potency of gene in 35 independent lung cancer cohorts.

MATERIALS AND METHODS

Collection of Lung Cancer Datasets

The lung cancer cohorts for OSluca with expression profiling and clinical follow-up data were collected from PubMed, TCGA,[1] and GEO[2] by searching the keywords: "lung" AND "cancer" AND "survival" (**Table 1**). The dataset for each cohort that met these following criteria will be included in OSluca: (1) have RNA sequencing or gene microarray data; (2) have complete follow-up data, such as overall survival and status (Liu et al., 2018); (3) all the data were specific for lung cancer, not from secondary or metastatic lung tumor from other types of tumors; (4) the cohort size is no less than 30 cases. The primary clinical pathological characteristics of lung cancer patients are listed in **Table 1**.

Construction of OSluca Web Server

Online consensus Survival for Lung Cancer is built in a tomcat server as previously described with minor modifications (Wang et al., 2019b,c; Xie et al., 2019b,c; Yan et al., 2019; Zhang et al., 2019). Briefly, front-end application was used for inputting query and displaying the results. Java and R package were used to analyze request and output the results. In addition, profiles and clinical information were stored in the SQL Server database. The prognostic significance of inputted gene is determined by analyzing the association of gene expression and survival time using the R package "survival." In addition, a genome-wide pre-calculation of Cox proportional regression for all the human genes were performed as well, and the home page of OSluca could display the survival analysis summary with a forest plot and a table of Cox proportional regression result for inputted gene in all cohorts with *P*-value and HR [(95% confidence interval (CI)] with the built-in upper 25% cutoff. The R package "forestplot" was used to produce the forest plot for inputted gene in OSluca web server.

[1] https://cancergenome.nih.gov/

[2] www.ncbi.nlm.nih.gov/geo/

Validation of Previously Reported Prognostic Biomarkers of Lung Cancer in OSluca

Keywords including "lung cancer," "survival," "biomarker," and "prognosis" were used to search biomarkers of lung cancer in NCBI PubMed. We finally obtained 104 prognostic biomarkers using the following criteria (**Table 2**): (1) immunohistochemistry (IHC) or qRT-PCR (qPCR) detection of biomarkers in primary cancer tissue; (2) a significant association between biomarker and survival; (3) the sample size must be above 50 cases; (4) the study was published in the English for full access.

Statistical Analysis

The association of lung cancer clinical factors and survival outcomes was analyzed by GraphPad Prism 8.0 software. The Cox proportional hazards regression and Kaplan Meier plot functions from R package "survival" were used in the OSluca to determine the association between gene expression and survival. The $P \leq 0.05$ was considered statistically significant.

RESULTS

Clinical Characteristics of Lung Cancer Patients in OSluca

To develop an online survival web server for lung cancer, we collected 35 published high-throughput profiling datasets of lung cancer with long-term follow-up information (2 TCGA datasets, 32 GEO datasets, and 1 Roepman dataset). TCGA comprises 513 lung adenocarcinoma cases and 499 squamous cell carcinoma cases (**Tables 1**, **2**). GEO cohorts and Roepman cohort had more than 4,000 samples and 172 samples, respectively, as shown in **Table 2**. 4,901 patients have OS (overall survival) data; 2,176 patients have DSS (disease-specific survival) data; and 2,075 patients have PFI (progression-free interval or recurrence-free survival) data, while 608 patients have DFI (disease-free interval) data. The results showed that the patients with lung adenocarcinoma significantly survive longer than those of other histological lung cancer, and small cell lung cancer is associated with the worst prognosis compared to other types of lung cancer (**Figure 1A**). Moreover, other clinical characteristics can also prominently affect patients' prognosis, such as gender ($P < 0.0001$), stage ($P < 0.0001$), p-TNM stage ($P < 0.0001$), and smoking status ($P < 0.0001$) (**Figures 1B–E**). Besides, these risk factors can influence other survival endpoints, such as PFI (data not shown). These results are in accordance with previous researches (Mao et al., 2016; Bray et al., 2018).

Construction and Usage of Prognostic Web Server OSluca

Online consensus Survival for Lung Cancer includes a set of optional clinico-pathological factors, such as age, sex, histological type, grade, smoking status, and so on. Four survival endpoints can be selected basing on original patient outcomes, containing OS, DSS, DFI, and PFI (Liu et al., 2018). In order to make the

TABLE 1 | Summary of clinical characteristics of lung cancer cohorts in Online Consensus Survival for Lung Cancer (OSluca).

	NSCLC					SCLC (*N* = 223)	#NA (*N* = 85)
	NSCLC, Total (*N* = 4937)	AD (*N* = 3345)	SCC (*N* = 1381)	LCC (*N* = 197)	NOS (*N* = 194)		
Age, year	64 (13–91)	64 (13–90)	66 (39–83)	63 (39–81)	62 (22–80)	64 (40–83)	58 (15–82)
Gender							
Male, %	52.6	46.9	68.3	77.2	12.9	58.1	50
Female, %	38.8	47.7	23.7	18.1	12.4	41.9	50
#NA, %	8.6	5.4	8.0	4.7	73.7	0	0
Stage*							
I, *n*	2301	1,653	567	66	28	10	9
II, *n*	889	500	347	27	15	5	4
III, *n*	595	366	199	18	12	2	3
IV, *n*	101	73	13	2	13	0	0
T stage 1/2/3/4/#NA	646/1074/230/103/2884	468/663/102/49/2063	155/362/109/39/716	20/44/17/9/107	3/5/2/6/178	11/13/5/4/190	28/20/10/6/21
N Stage 0/1/2/3/#NA	1638/495/280/21	1038/254/198/5/1859	549/218/70/7/537	48/20/17/5/107	3/3/4/4/180	14/4/12/6/187	33/25/5/1/21
M stage 0/1/#NA	1685/42/3210	853/26/2466	740/8/633	82/2/113	10/6/178	33/4/186	63/2/20
Smoking/non-smoking/#NA	1839/262/2836	1112/256/1977	618/3/760	40/1/156	9/2/183	18/1/204	9/8/68
OS, mo	46 (0.03–256)	48 (0.03–242)	41 (0.03–256)	46 (0.1–216)	38 (0.5–208)	51 (2–211)	68 (2–244)
DSS, mo	42 (0.03–256)	43 (0.19–242)	41 (0.03–256)	45 (1–216)	36 (6–76)	24 (2–140)	69 (2–244)
DFI, mo	33 (0.16–242)	32 (0.6–242)	34 (0.16–159)	–	–	–	–
PFI, mo	33 (0.03–242)	36 (0.03–242)	30 (0.03–180)	53 (1.8–164)	4 (0.23–54)	–	30 (2–73)

*NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer; AD, adenocarcinoma; SCC, squamous cell carcinoma; LCC, large cell cancer; NOS, NSCLC, not otherwise specified; F, female; M, male; n, number; mo, months; OS, overall survival; DSS, disease-specific survival; DFI, disease-free interval; PFI, progression-free interval or recurrence free survival. *The stage only counts stages of lung cancer patients described in the original datasets; #NA, data lost or unknown.*

user clearly see the prognostic effect of interested gene, a meta-analysis is to summarize the prognostic value for each gene on the home page of OSluca. Briefly, after the user types the official gene symbol into the input box on the home page, OSluca will display the survival analysis summary with a forest plot and a table from Cox proportional regression in each cohort and combined cohorts (combining all the datasets together). Take the tumor suppressor gene *TP53* (tumor protein p53) as an example and type "TP53" into the gene symbol box and click on "Survival analysis" (**Figure 2A**, left). The meta-analysis results with a forest plot and a survival table for the *TP53* gene, will display the *P*-value and HR with 95% CI of each cohort and the combined cohorts (**Figure 2A**, right). Then, the user can easily obtain KM plots of separate cohorts such as GSE30219 dataset by clicking on the "Go" button in the survival table (**Figure 2B**). In addition, it is also available to use a subgroup of certain cohort to obtain specific prognostic information with selectable risk factors, such as cutoff value, histological type, grade, etc. Briefly, OSluca can output survival rates displaying a forest plot and a survival table with KM plot and *P*-value to measure the association between the investigated gene and survival rate.

Validation of Previously Reported Lung Cancer Prognostic Biomarkers in OSluca

A search for lung cancer biomarkers was performed using a set of keywords in NCBI PubMed, including "lung cancer," "survival," "biomarker," and "prognosis." In total, we collected 104 published lung cancer prognostic biomarkers verified by IHC or qPCR (**Supplementary Table S1**) to evaluate the performance of OSluca. For example, Hsu et al. reported that *ERO1L* (ERO1-like protein alpha, also named *ERO1A*) is significantly overexpressed in tumor tissue and could be as a poor prognostic biomarker for lung adenocarcinoma (Hsu et al., 2016). The prognostic analysis of *ERO1L* in OSluca showed that high expression of *ERO1L* gene is significantly associated with poor outcome in eight out of nine cohorts (Top 9 cohorts, the sample size above 150 cases) (**Figures 3A–H**), except the Roepman dataset (**Figure 3I**). Next, each published biomarker was investigated in the Top 9 cohorts in OSluca, and the results showed that approximately 66% of biomarkers (69/104) were consistent with original published findings (**Supplementary Table S1**). Meanwhile, OSluca can be used to perform the outcome meta-analysis of the interested gene that showed that 14% (14/104) (**Supplementary Table S1**) of published prognostic genes have the similar prognostic values in one or multiple OSluca cohorts as reported in the literature, but these genes also showed the opposite outcomes in some other cohorts from OSluca. These genes need further investigations, such as the *DDIT3* gene (**Supplementary Figure S2** and **Supplementary Table S1**). In contrast, there are some prognostic biomarkers, which have been shown different outcomes between OSluca and previous findings. A total of 9% of the published prognostic genes showed opposite outcome results between OSluca and literatures (9/104) (see **Supplementary Table S1**), suggesting that these genes need further validation. For example, the transcription factor *KLF15* (Krüppel-like factor 15) had been proven to be higher in tumor tissue than that of adjacent non-tumor tissue and played

TABLE 2 | Clinico-pathological traits of lung cancer cohorts.

Datasets	Cohorts	Platform	Histological type	Survival	Samples	References
Rockville	GSE102287	GPL570	AD/SCC/NOS	OS	32	Mitchell et al., 2017
Heidelberg	GSE10245	GPL570	AD/SCC	OS	58	Kuner et al., 2009
Koto-ku	GSE1037	GPL962	AD/SCC/SCLC	OS	61	Jones et al., 2004
Basel	GSE11117	GPL6650	AD/SCC/NOS	OS	41	Baty et al., 2010
Nagoya	GSE11969	GPL7015	AD/SCC/LCC	OS	149	Takeuchi et al., 2006
Groningen	GSE12428	GPL1708	SCC	OS	34	Boelens et al., 2009
Nagoya	GSE13213	GPL6480	AD	OS	117	Tomida et al., 2009
Toronto	GSE14814	GPL96	AD/SCC /NOS	OS/DSS	133	Zhu et al., 2010
Chapel Hill	GSE17710	GPL9053	SCC	OS/PFI	56	Wilkerson et al., 2010
Rotterdam	GSE19188	GPL570	AD/SCC/LCC	OS	82	Hou et al., 2010
Chapel Hill	GSE26939	GPL9053	AD	OS	116	Wilkerson et al., 2012
Dallas	GSE29013	GPL570	AD/SCC	OS/PFI	55	Xie et al., 2011
Lund	GSE29066	GPL6947	AD/SCC/SCLC	OS	68	Staaf et al., 2012, 2013
La Tronche	GSE30219	GPL570	AD/SCC/SCLC/LCC	OS/DFS	293	Rousseaux et al., 2013
Chuo-ku	GSE31210	GPL570	AD	OS /PFI	226	Okayama et al., 2012
Durham	GSE3141	GPL570	AD/SCC	OS	111	Bild et al., 2006
Dallas	GSE31908	GPL96/97	AD	OS	30	NA
Houston	GSE33072	GPL6244	AD/SCC	PFI	66	Byers et al., 2013
Uppsala	GSE37745	GPL570	AD/SCC/LCC	PFI	196	Botling et al., 2013
Dallas	GSE41271	GPL6884	AD/SCC/LCC	OS/PFI	275	Sato et al., 2013
San Diego	GSE4573	GPL96	SCC	OS	130	Raponi et al., 2006
Nagoya	GSE4716	GPL3696/3694	AD/SCC/LCC	OS	50	Tomida et al., 2004
Toronto	GSE50081	GPL570	AD/SCC/LCC	OS/DFS	181	Der et al., 2014
Brisbane	GSE5123	GPL3877	SCC	OS	51	Larsen et al., 2007b
Brisbane	GSE5828	GPL3877	SCC	OS	59	Larsen et al., 2007a
Brisbane	GSE5843	GPL3877	AD	OS	48	Larsen et al., 2007c
St. Louis	GSE6253	GPL8300	AD/SCC/NOS	DSS	34	Lu et al., 2006
Bethesda	GSE63459	GPL6883	AD	OS	33	Robles et al., 2015
Stanford	GSE67639	GPL570	AD/SCC/NOS	OS	1106	Gentles et al., 2015
Rockville	GSE68465	GPL96	AD	OS/PFI	442/363	Shedden et al., 2008
Rockville	GSE68571	GPL80	AD	OS	86	Beer et al., 2002
Seoul	GSE8894	GPL570	AD/SCC	PFI	138	Lee et al., 2008
NIH and NHGRI	TCGA	DCC	AD	OS/DSS/DFI/PFI	513/478/306/513	The Cancer Genome Atlas Research Network, 2014; Liu et al., 2018
NIH and NHGRI	TCGA	DCC	SCC	OS/DSS/DFI/PFI	498/452/303/499	Hammerman et al., 2012; The Cancer Genome Atlas Research Network, 2012; Liu et al., 2018
Reopman	Roepman		AD/SCC/LCC/NOS	OS	172	Roepman et al., 2009

NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer; AD, adenocarcinoma; SCC, squamous cell carcinoma; LCC, large cell carcinoma; NOS, not otherwise specified; OS, overall survival; DSS, disease-specific survival; DFI, disease-free interval; PFI, progression-free interval.

an important role in promoting proliferation and carcinoma diversification in lung adenocarcinoma, associated with poor prognostic outcome (Gao et al., 2017). It was not anticipated that the patients with high expression of *KLF15* have better survival than those with low expression (**Supplementary Table S1** and **Supplementary Figure S1**). The OSluca result for the *KLF15* gene was consistent with other prognostic analysis tools (Győrffy et al., 2013; Anaya, 2016), such as the KM plotter [$P < 0.001$, HR (95% CI) = 0.4 (0.28–0.58)]. In addition, the remaining 12 of 104 previously published prognostic biomarkers (11%) were not significant for prognostic analysis in the Top 9 cohorts in OSluca, but 8 of them (8/12) are significant in one or multiple datasets other than the Top 9 cohorts in OSluca (data not shown). All in all, the OSluca server is an interactive and free web server for researchers to develop potential prognostic biomarkers for lung cancer.

DISCUSSION

Owing to tumor molecular heterogeneity, the prognosis of lung cancer patients is variable and difficult to predict. The prognosis of patients suffering from lung cancer had been demonstrated to be highly dependent on clinical factors

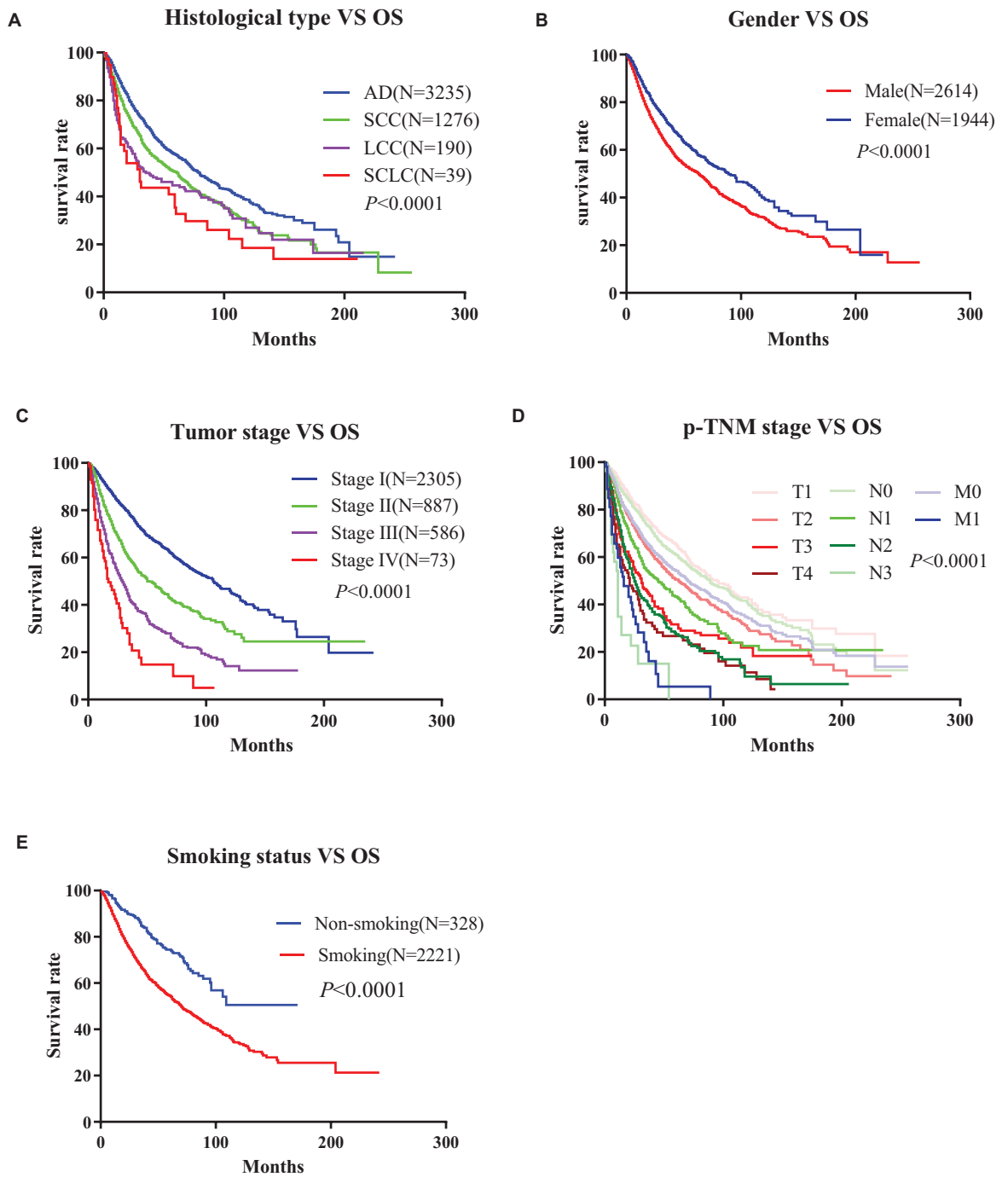


FIGURE 1 | Correlation between the clinico-pathologic characteristics and overall survival of lung cancer in Online Consensus Survival for Lung Cancer (OSluca). **(A)** Correlation between histological types and OS. **(B)** Correlation between gender and OS. **(C)** Correlation between tumor stages and OS. **(D)** Correlation between p-TNM stages and OS. **(E)** Correlation between smoking status and OS. OS, overall survival; AD, adenocarcinoma; SCC, squamous cell carcinoma; LCC, large cell cancer.

of the patient, such as histological type, smoking status, and so on. However, it is also an imperative need to exploit novel prognostic biomarkers for determining the risk of cancerous lesions and predicting lung cancer patient outcomes by all available means, especially by high-throughput sequencing technologies. However, one major challenge to non-bioinformatics researchers is how to integrate the high-dimension profiling datasets of lung cancer and discover new biomarkers to potentially guide prognostic stratification. Previous studies had revealed that the online prognostic web

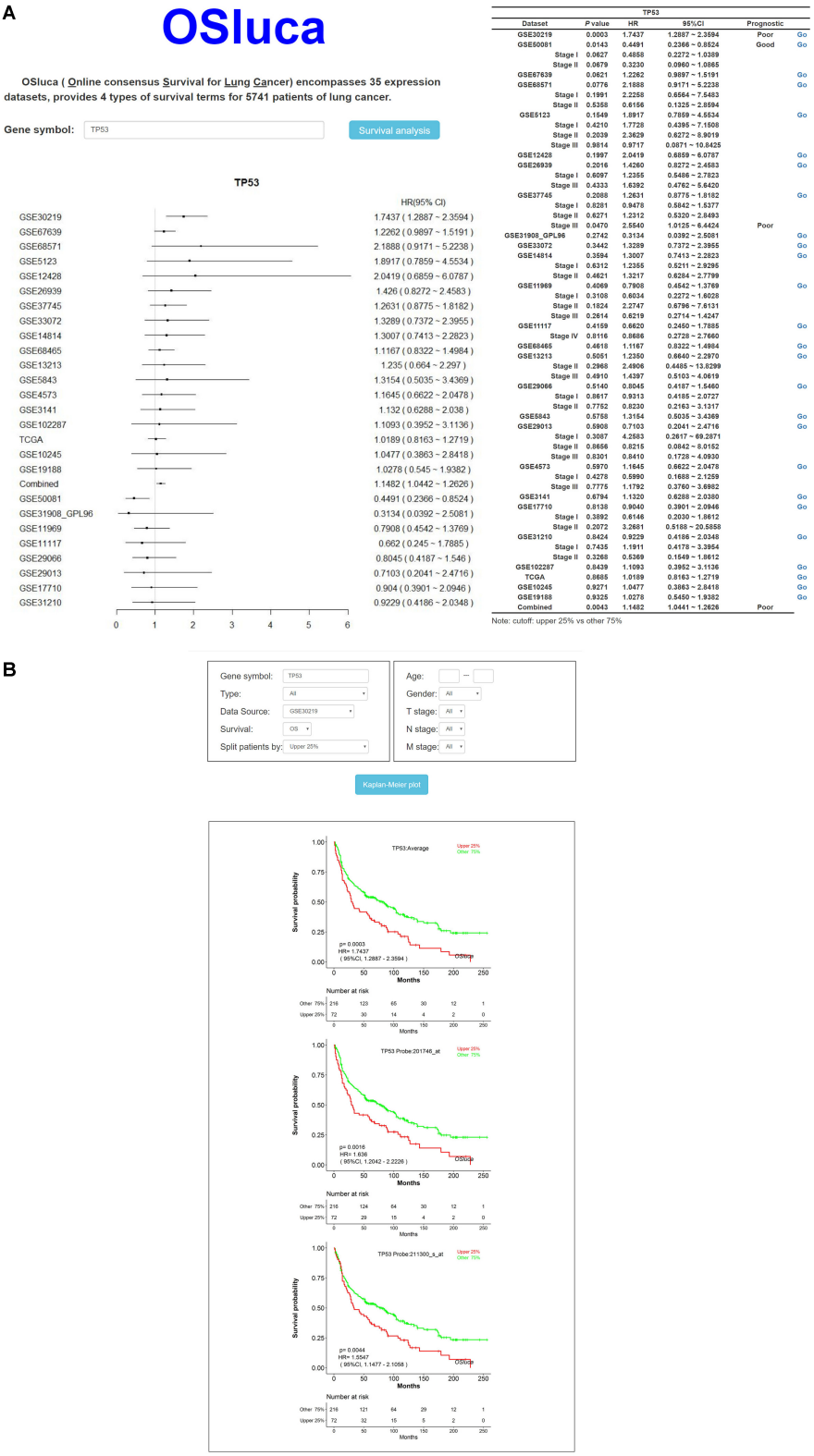


FIGURE 2 | The output home page and KM output web subpage in OSluca for lung cancer. **(A)** Home page of OSluca with *TP53* gene survival analysis, containing prognostic meta-analysis of a forest plot and a survival table. **(B)** KM plots of *TP53* gene in the GSE30219 cohort. Note: the cutoff value is the upper 25% vs. other 75%. The "Combined" in forest plot and survival table means the overall prognostic significance of inputted gene in a pooling cohort with all the datasets. TP53, tumor protein p53.

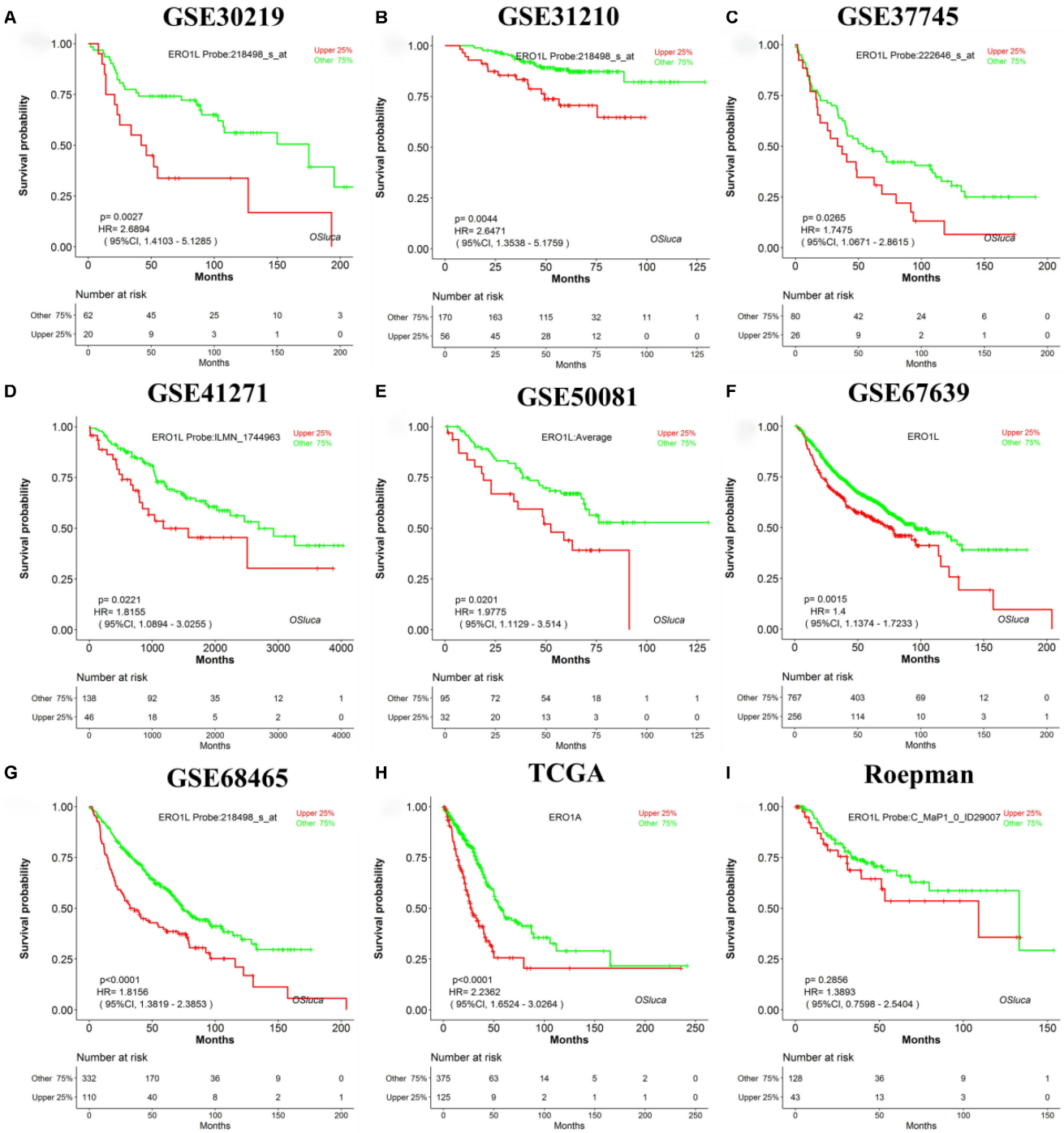


FIGURE 3 | Validation of a previously reported biomarker *ERO1L* in OSluca. Overexpression of *ERO1L* in tumor tissue is suggested as a worse survival biomarker in lung adenocarcinoma. **(A)** Overall survival (OS) of *ERO1L* gene in GSE30219 cohort. **(B)** OS in GSE31210 cohort. **(C)** OS in GSE37745 cohort. **(D)** OS in GSE41271 cohort. **(E)** OS in GSE50081 cohort. **(F)** OS in GSE67639 cohort. **(G)** OS in GSE68465 cohort. **(H)** OS in TCGA in lung adenocarcinoma. **(I)** OS in Roepman cohort. The histological type of all the above cohorts is lung adenocarcinoma. ERO1L, ERO1-like protein alpha (also named ERO1A).

servers of cancer (Elfilali et al., 2006; Mizuno et al., 2009; Goswami and Nakshatri, 2013; Győrffy et al., 2013; Tang et al., 2017) could substantially help researchers to discover potential biomarkers (Zheng et al., 2020). Herein, we developed a free web server OSluca to assess the prognostic value of the interesting gene in multiple cohorts of lung cancers. In OSluca, all the lung cancer cases are originated from the organ lung, not the second cancer from other cancers or

organs. As a result, the prognostic specificity is only for lung cancer. Nevertheless, its prognostic significance in other types of cancers is also worth to be determined. To access the repeatability of previously reported prognostic biomarkers in OSluca, we collected 104 previously published prognostic biomarkers of lung cancer identified by qPCR or IHC, and tested their prognostic significance in OSluca. The testing results showed that most of the biomarkers were verified in OSluca and were confirmed for the published findings. Nevertheless, some genes showed different prognostic outcomes compared to previous literatures.

The advantage of OSluca over other online prognostic web servers is that the size of lung cancer samples in OSluca is large, and tens of independent cohorts are available, which is extremely valuable for the identification and validation of cancer prognostic biomarkers, since the most important part for the biomarker development is independent validation across different datasets/cohorts. The limitation of the current study is that OSluca can only test a single gene for outcome analysis. In summary, OSluca is a free web server for non-bioinformatics researchers to study potential lung cancer prognostic biomarkers, accessed at http://bioinfo.henu.edu.cn/LUCA/LUCAList.jsp.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the TCGA, NCBI GEO, and Roepman dataset.

AUTHOR CONTRIBUTIONS

XG: research design. QW and XG: establish OSluca web server. ZY, ZL, and XS: deal with RNA sequencing with clinical data of lung cancer. ZY, LX, XS, LZ, YL, and XG: draft of the manuscript. YD, XS, LZ, PS, YL, TX, and JM: collect previously reported biomarkers of lung cancer. ZY, LX, LZ, WZ, YZ, and XG: critical revision of the manuscript.

FUNDING

This study was supported by the following funding: The Kaifeng Science and Technology Major Project (18ZD008), the National Natural Science Foundation of China (Nos. 81602362 and 81801569), the Program for Science and Technology Development in Henan Province (Nos. 162102310391, 172102210187, and 192102310302), the Program for Young Key Teacher of Henan Province (2016GGJS-214), the supporting grants of Henan University (Nos. 2015YBZR048 and B2015151), and the Yellow River Scholar Program (No. H2016012).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00420/full#supplementary-material

REFERENCES

Anaya, J. (2016). OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peerj Comput. Sci.* 2:e67.

Baty, F., Facompre, M., Kaiser, S., Schumacher, M., Pless, M., Bubendorf, L., et al. (2010). Gene profiling of clinical routine biopsies and prediction of survival in non-small cell lung cancer. *Am. J. Respir. Crit. Care Med.* 181, 181–188. doi: 10.1164/rccm.200812-1807OC

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824. doi: 10.1038/nm733

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Boelens, M. C., van den Berg, A., Fehrmann, R. S., Geerlings, M., de Jong, W. K., te Meerman, G. J., et al. (2009). Current smoking-specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer. *J. Pathol.* 218, 182–191. doi: 10.1002/path.2520

Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., et al. (2013). Biomarker discovery in Non–small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.* 19, 194–204. doi: 10.1158/1078-0432.ccr-12-1139

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Byers, L. A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., et al. (2013). An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* 19, 279–290. doi: 10.1158/1078-0432.CCR-12-1558

Der, S. D., Sykes, J., Pintilie, M., Zhu, C.-Q., Strumpf, D., Liu, N., et al. (2014). Validation of a histology-independent prognostic gene signature for early-stage, non–small-cell lung cancer including stage IA patients. *J. Thorac. Oncol.* 9, 59–64. doi: 10.1097/JTO.0000000000000042

Dong, H., Wang, Q., Zhang, G., Li, N., Yang, M., An, Y., et al. (2020). OSdlbcl: an online consensus survival analysis web server based on gene expression profiles of diffuse large B-cell lymphoma. *Cancer Med.* 9, 1790–1797. doi: 10.1002/cam4.2829

Elfilali, A., Lair, S., Verbeke, C., La Rosa, P., Radvanyi, F., and Barillot, E. (2006). ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. *Nucleic Acids Res.* 34, D613–D616. doi: 10.1093/nar/gkj022

Gao, L., Qiu, H., Liu, J., Ma, Y., Feng, J., Qian, L., et al. (2017). KLF15 promotes the proliferation and metastasis of lung adenocarcinoma cells and has potential as a cancer prognostic marker. *Oncotarget* 8, 109952–109961. doi: 10.18632/oncotarget.21972

Gentles, A. J., Bratman, S. V., Lee, L. J., Harris, J. P., Feng, W., Nair, R. V., et al. (2015). Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage Non-small cell lung cancer. *J. Natl. Inst.* 107:djv211. doi: 10.1093/jnci/djv211

Goswami, C. P., and Nakshatri, H. (2013). PROGgene: gene expression based survival analysis web application for multiple cancers. *J. Clin. Bioinform.* 3:22. doi: 10.1186/2043-9113-3-22

Győrffy, B., Surowiak, P., Budczies, J., and Lánczky, A. (2013). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One* 8:e82241. doi: 10.1371/journal.pone.0082241

Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519–525. doi: 10.1038/nature11404

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors

from 33 Types of cancer. *Cell* 173, 291.e6–304.e6. doi: 10.1016/j.cell.2018.03.022

Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS One* 5:e10312. doi: 10.1371/journal.pone.0010312

Hsu, C.-H., Hsu, C.-W., Hsueh, C., Wang, C.-L., Wu, Y.-C., Wu, C.-C., et al. (2016). Identification and characterization of potential biomarkers by quantitative tissue proteomics of primary lung adenocarcinoma. *Mol. Cell. Proteomics* 15, 2396–2410. doi: 10.1074/mcp.M115.057026

Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., et al. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* 363, 775–781. doi: 10.1016/s0140-6736(04)15693-6

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., et al. (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* 63, 32–38. doi: 10.1016/j.lungcan.2008.03.033

Larsen, J. E., Pavey, S. J., Bowman, R., Yang, I. A., Clarke, B. E., Colosimo, M. L., et al. (2007a). Gene expression of lung squamous cell carcinoma reflects mode of lymph node involvement. *Eur. Respir. J.* 30, 21–25. doi: 10.1183/09031936.00161306

Larsen, J. E., Pavey, S. J., Passmore, L. H., Bowman, R., Clarke, B. E., Hayward, N. K., et al. (2007b). Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis* 28, 760–766. doi: 10.1093/carcin/bgl207

Larsen, J. E., Pavey, S. J., Passmore, L. H., Bowman, R. V., Hayward, N. K., and Fong, K. M. (2007c). Gene expression signature predicts recurrence in lung adenocarcinoma. *Clin. Cancer Res.* 13, 2946–2954. doi: 10.1158/1078-0432.ccr-06-2525

Lee, E.-S., Son, D.-S., Kim, S.-H., Lee, J., Jo, J., Han, J., et al. (2008). Prediction of recurrence-free survival in postoperative Non–small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin. Cancer Res.* 14, 7397–7404. doi: 10.1158/1078-0432.ccr-07-4937

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400.e11–416.e11. doi: 10.1016/j.cell.2018.02.052

Lu, Y., Lemon, W., Liu, P.-Y., Yi, Y., Morrison, C., Yang, P., et al. (2006). A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med.* 3:e030467. doi: 10.1371/journal.pmed.0030467

Mao, Y., Yang, D., He, J., and Krasna, M. J. (2016). Epidemiology of lung cancer. *Surg. Oncol. Clin. North Am.* 25, 439–445. doi: 10.1016/j.soc.2016.02.001

Meyerson, M., and Carbone, D. (2005). Genomic and proteomic profiling of lung cancers: lung cancer classification in the age of targeted therapy. *J. Clin. Oncol.* 23, 3219–3226. doi: 10.1200/JCO.2005.15.511

Mitchell, K. A., Zingone, A., Toulabi, L., Boeckelman, J., and Ryan, B. M. (2017). Comparative transcriptome profiling reveals coding and noncoding RNA differences in NSCLC from african americans and european americans. *Clin. Res.* 23, 7412–7425. doi: 10.1158/1078-0432.ccr-17-0527

Mizuno, H., Kitada, K., Nakai, K., and Sarai, A. (2009). PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med. Genomics* 2:18. doi: 10.1186/1755-8794-2-18

Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., et al. (2012). Identification of genes up-regulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.* 72, 100–111. doi: 10.1158/0008-5472.can-11-1403

Ramalingam, S. S., Owonikoko, T. K., and Khuri, F. R. (2011). Lung cancer: new biological insights and recent therapeutic advances. *Cancer J. Clin.* 61, 91–112. doi: 10.3322/caac.20102

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., et al. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.* 66, 7466–7472. doi: 10.1158/0008-5472.can-06-1191

Robles, A. I., Arai, E., Mathé, E. A., Okayama, H., Schetter, A. J., Brown, D., et al. (2015). An integrated prognostic classifier for stage I lung adenocarcinoma based on mRNA, microRNA, and DNA methylation biomarkers. *J. Thora.c Oncol.* 10, 1037–1048. doi: 10.1097/JTO.0000000000000560

Roepman, P., Jassem, J., Smit, E. F., Muley, T., Niklinski, J., van de Velde, T., et al. (2009). An Immune response enriched 72-gene prognostic profile for early-stage non–small-cell lung cancer. *Clin. Cancer Res.* 15, 284–290. doi: 10.1158/1078-0432.ccr-08-1258

Rousseaux, S., Debernardi, A., Jacquiau, B., Vitte, A.-L., Vesin, A., Nagy-Mignotte, H., et al. (2013). Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* 5:ra66. doi: 10.1126/scitranslmed.3005723

Sato, M., Larsen, J. E., Lee, W., Sun, H., Shames, D. S., Dalvi, M. P., et al. (2013). Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol. Cancer Res.* 11, 638–650. doi: 10.1158/1541-7786.MCR-12-0634-T

Shedden, K., Taylor, J. M. G., Enkemann, S. A., Tsao, M.-S., Yeatman, T. J., Gerald, W. L., et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* 14, 822–827. doi: 10.1038/nm.1790

Staaf, J., Isaksson, S., Karlsson, A., Jonsson, M., Johansson, L., Jonsson, P., et al. (2013). Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. *Int. J. Cancer* 132, 2020–2031. doi: 10.1002/ijc.27879

Staaf, J., Jönsson, G., Jönsson, M., Karlsson, A., Isaksson, S., Salomonsson, A., et al. (2012). Relation between smoking history and gene expression profiles in lung adenocarcinomas. *BMC Med. Genomics* 5:22. doi: 10.1186/1755-8794-5-22

Takeuchi, T., Tomida, S., Yatabe, Y., Kosaka, T., Osada, H., Yanagisawa, K., et al. (2006). Expression profile–defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J. Clin. Oncol.* 24, 1679–1688. doi: 10.1200/JCO.2005.03.8224

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

The Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519–525.

The Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/s41586-018-0228-6

Tomida, S., Koshikawa, K., Yatabe, Y., Harano, T., Ogura, N., Mitsudomi, T., et al. (2004). Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* 23, 5360–5370. doi: 10.1038/sj.onc.1207697

Tomida, S., Takeuchi, T., Shimada, Y., Arima, C., Matsuo, K., Mitsudomi, T., et al. (2009). Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. *J. Clin. Oncol.* 27, 2793–2799. doi: 10.1200/JCO.2008.19.7053

Wang, F., Wang, Q., Li, N., Ge, L., Yang, M., An, Y., et al. (2019a). OSuvm: an interactive online consensus survival tool for uveal melanoma prognosis analysis. *Mol. Carcinog.* 59, 56–61. doi: 10.1002/mc.23128

Wang, F., Wang, Q., Li, N., Ge, L., Yang, M., An, Y., et al. (2020). OSuvm: an interactive online consensus survival tool for uveal melanoma prognosis analysis. *Mol. Carcinog.* 59, 56–61.

Wang, Q., Xie, L., Dang, Y., Sun, X., Xie, T., Guo, J., et al. (2019b). OSlms: a web server to evaluate the prognostic value of genes in leiomyosarcoma. *Front. Oncol.* 9:190. doi: 10.3389/fonc.2019.00190

Wang, Q., Zhang, L., Yan, Z., and Xie, L. (2019c). OScc: an online survival analysis web server to evaluate the prognostic value of biomarkers in cervical cancer. *Future Oncol.* 15, 3693–3699. doi: 10.2217/fon-2019-0412

Wilkerson, M. D., Yin, X., Hoadley, K. A., Liu, Y., Hayward, M. C., Cabanski, C. R., et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* 16, 4864–4875. doi: 10.1158/1078-0432.CCR-10-0199

Wilkerson, M. D., Yin, X., Walter, V., Zhao, N., Cabanski, C. R., Hayward, M. C., et al. (2012). Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PloS One* 7:e36530. doi: 10.1371/journal.pone.0036530

Xie, L., Dang, Y., Guo, J., Sun, X., Xie, T., Zhang, L., et al. (2019a). High KRT8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes* 10:36. doi: 10.3390/genes10010036

Xie, L., Wang, Q., Dang, Y., Ge, L., Sun, X., Li, N., et al. (2019b). OSkirc: a web tool for identifying prognostic biomarkers in kidney renal clear cell carcinoma. *Future Oncol.* 15, 3103–3110. doi: 10.3892/ol.2019.10440

Xie, L., Wang, Q., Nan, F., Ge, L., and Dang, Y. (2019c). OSacc: gene expression-based survival analysis web tool for adrenocortical carcinoma. *Cancer Manag Res.* 11, 9145–9152. doi: 10.2147/cmar.s215586

Xie, Y., Xiao, G., Coombes, K. R., Behrens, C., Solis, L. M., Raso, G., et al. (2011). Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin. Cancer Res.* 17, 5705–5714. doi: 10.1158/1078-0432.CCR-11-0196

Yan, Z., Wang, Q., Sun, X., Ban, B., Lu, Z., Dang, Y., et al. (2019). OSbrca: a web server for breast cancer prognostic biomarker investigation with massive data from tens of cohorts. *Front. Oncol.* 9:1349. doi: 10.3389/fonc.2019.01349

Zhang, G., Wang, Q., Yang, M., Yao, X., Qi, X., An, Y., et al. (2020). OSpaad: an online tool to perform survival analysis by integrating gene expression profiling and long-term follow-up data of 1319 pancreatic carcinoma patients. *Mol. Carcinog.* 59, 304–310. doi: 10.1002/mc.23154

Zhang, G., Wang, Q., Yang, M., Yuan, Q., Dang, Y., Sun, X., et al. (2019). OSblca: a web server for investigating prognostic biomarkers of bladder cancer patients. *Front. Oncol.* 9:466. doi: 10.2217/fon-2019-0296

Zheng, H., Zhang, G., Zhang, L., Wang, Q., Li, H., Han, Y., et al. (2020). Comprehensive review of web servers and bioinformatics tools for cancer prognosis analysis. *Front. Oncol.* 10:68. doi: 10.3389/fonc.2020.00068

Zhu, C.-Q., Ding, K., Strumpf, D., Weir, B. A., Meyerson, M., Pennell, N., et al. (2010). Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J.Clin. Oncol.* 28, 4417–4424. doi: 10.1200/JCO.2009.26.4325

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 07 July 2020
doi: 10.3389/fonc.2020.00962

Single-Nucleotide Polymorphism Array Technique Generating Valuable Risk-Stratification Information for Patients With Myelodysplastic Syndromes

Xia Xiao[1], Xiaoyuan He[2], Qing Li[1], Wei Zhang[3], Haibo Zhu[1], Weihong Yang[4], Yuming Li[1], Li Geng[1], Hui Liu[3], Lijuan Li[3], Huaquan Wang[3], Rong Fu[3], Mingfeng Zhao[1,2†], Zhong Chen[4*†] and Zonghong Shao[3*†]*

[1] Department of Hematology, Tianjin First Central Hospital, Tianjin, China, [2] Department of Clinical Medicine, Nankai University School of Medicine, Tianjin, China, [3] Department of Hematology, Tianjin Medical University General Hospital, Tianjin, China, [4] Wuhan Kindstar Diagnostics Co./Kindstar Global Gene (Beijing) Technology, Inc., Wuhan, China

OPEN ACCESS

Edited by:
Liuyang Wang,
Duke University, United States

Reviewed by:
Lina Shao,
University of Michigan, United States
Giovana Tardin Torrezan,
A.C.Camargo Cancer Center, Brazil

***Correspondence:**
Mingfeng Zhao
mingfengzhao@sina.com
Zhong Chen
chenzhong@kindstar.com.cn
Zonghong Shao
shaozonghong@sina.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 23 October 2019
Accepted: 15 May 2020
Published: 07 July 2020

Citation:
Xiao X, He X, Li Q, Zhang W, Zhu H, Yang W, Li Y, Geng L, Liu H, Li L, Wang H, Fu R, Zhao M, Chen Z and Shao Z (2020) Single-Nucleotide Polymorphism Array Technique Generating Valuable Risk-Stratification Information for Patients With Myelodysplastic Syndromes. Front. Oncol. 10:962. doi: 10.3389/fonc.2020.00962

Background: Chromosomal abnormalities play an important role in the diagnosis and prognosis of patients with myelodysplastic syndromes (MDSs). The single-nucleotide polymorphism array (SNP-A) technique has gained popularity due to its improved resolution compared to that of metaphase cytogenetic (MC) analysis.

Methods: A total of 376 individuals were recruited from two medical centers in China, including 350 patients and 26 healthy individuals. Among these patients, 200 were diagnosed with *de novo* MDS, 25 with myeloproliferative neoplasm (MPN), 63 with primary acute myeloid leukemia (AML), and 62 with idiopathic cytopenia of undetermined significance (ICUS). We evaluated the significance of abnormal chromosomes detected by SNP-A in the diagnosis and prognosis of MDS-related disorders.

Results: (1) When certain chromosomal abnormalities could not be detected by conventional MC methods, these abnormalities could be detected more efficiently by the SNP-A method. With SNP-A, the detection rates of submicroscopic or cryptic aberrations in the MDS, MPN, and AML patients with normal MC findings were 32.8, 30.8, and 30%, respectively. (2) The chromosomal abnormalities detected by SNP-A had a very important value for the prognosis of patients with MDSs, especially in the low-risk group. The survival of patients with abnormal chromosomes detected by SNP-A was significantly lower than that of patients with no detected chromosomal abnormalities; this difference was observed in overall survival (OS) ($P = 0.001$) and progression-free survival (PFS) [24 months vs. not reach (NR); $P = 0.008$]. The patients with multiple chromosomal abnormalities detected by SNP-A had an inferior prognosis, and SNP-A abnormalities (≥3 per patient) were found to be an independent predictor of poor prognosis in patients with MDSs [hazard ratio (HR) = 2.40, $P = 0.002$]. (3) Patients with ICUS may progress to myeloid malignancies, but most patients often maintain a stable ICUS status for many years without progression. An ICUS patient found to

have an MDS-related karyotype would be rediagnosed with MDS. SNP-A can efficiently detect chromosomal abnormalities, which would be important for assessing the evolution of ICUS. In our study, 17 ICUS patients with SNP-A-detected abnormalities developed typical MDSs.

Conclusions: SNP-A can help evaluate the prognosis of patients with MDSs and better assess the risk of disease progression for patients with ICUS.

Keywords: myelodysplastic syndrome (MDS), idiopathic cytopenia of undetermined significance (ICUS), single-nucleotide polymorphism (SNP), chromosome aberrations, prognosis

INTRODUCTION

Myelodysplastic syndromes (MDSs) are a heterogeneous group of malignant hematopoietic disorders characterized by dysplastic changes in one or more cell lineages, ineffective hematopoiesis, and a variable predilection to the development of acute myeloid leukemia (AML) (1). Karyotype analysis provides useful diagnostic and prognostic information for many hematological malignancies. Some chromosomal lesions have a significant impact on the prognosis of MDS patients, and poor chromosomal lesions significantly affect the survival of patients (2–4). In the prognostic algorithm and the Revised International Prognostic Scoring System (IPSS-R) of MDSs, cytogenetic results account for an important proportion. In addition, recent studies have shown that MDS patients with certain cytogenetic abnormalities may benefit from targeted therapies (5, 6). However, the standard metaphase cytogenetic (MC) technique, in general, can only detect chromosomal rearrangements of more than 10 Mb in size. Furthermore, chromosome banding analysis is dependent on the cell proliferation of MDS clones in culture to obtain metaphases. Thus, the MC technique will miss many important chromosome abnormalities, resulting in genomic aberrations detectable in only 40–50% of MDS patients (7, 8). Notably, ∼75–90% of chromosomal changes identified in MDSs are unbalanced aberrations, leading to gains or losses in all, or part, of specific chromosomes (3, 9, 10).

The single-nucleotide polymorphism array (SNP-A) technology relies on oligonucleotide probes corresponding to variants of the selected SNP allele. This method does not rely on cell division, has excellent resolution for unbalanced rearrangements, and overcomes some of the shortcomings of MC analysis. Since SNP-A has a higher analytical resolution than MC, SNP-A can detect submicroscopic or cryptic deletions or duplications. Another major advantage of SNP-A technology is its ability to recognize the loss of heterozygosity (LOH), which occurs when there is no simultaneous change in DNA copy number (CN), i.e., CN-neutral loss of heterozygosity. This defect is consistent with uniparental disomy (UPD). Acquired segmental UPD is increasingly recognized for its role in various tumors (11, 12). SNP-A-based genomic analysis has been applied in patients with various hematologic malignancies (2–4, 13, 14). A particularly interesting study by Mohamedali et al. (13) analyzed patients with low-risk MDS and found that 10% of these patients had a cryptic or submicroscopic deletion or duplication and 8% had gains. However, in general, the clinical significance of SNP-A-based analysis has not been fully realized.

The present study is aimed at developing a rational diagnostic algorithm for the detection of SNP-A-based genomic aberrations (unbalanced chromosome rearrangements and acquired UPDs) and establishing their clinical correlations in patients with MDS-related disorders. Based on the technical advantages of SNP-A, we assessed 376 cases of MDSs, various other myeloid disorders, and normal individuals. Our study represents the first such investigation in a large cohort of Chinese patients.

MATERIALS AND METHODS

Patients

A total of 376 individuals were recruited from the Department of Hematology at Tianjin Medical University General Hospital and Tianjin First Central Hospital from April 2013 to September 2016. These individuals included 200 patients with *de novo* MDS, 25 with myeloproliferative neoplasm (MPN), 63 with primary AML, and 62 with idiopathic cytopenia of undetermined significance (ICUS) as well as 26 healthy individuals. The 62 ICUS patients were initially suspected of having MDS but were subsequently redefined as having ICUS due to lack of typical abnormal karyotypes and morphological dysplasia as well as a proportion of blast cells <5% (10, 15). The MPN and AML cases served as the positive controls, and the healthy individuals served as the normal controls for the purposes of assay validation (**Table 1**).

Clinical data used for the assessment included age, sex, blood cell counts, bone marrow morphology, blast counts, and survival times, including progression-free survival (PFS) and overall survival (OS), for all patients (**Table 1**). The diagnosis and classification of MDS were in accordance with the Vienna diagnosis standard and the 2008 WHO classification (10, 16). Among the 200 MDS patients, 115 were males and 85 were females, aged from 12 to 87 years old with a median age of 60 years. According to the 2008 WHO classification standard (17), 10 cases were classified as refractory anemia with ringed sideroblasts (RARS), 34 as refractory cytopenia with unilineage dysplasia (RCUD), 68 as refractory cytopenia with multilineage dysplasia (RCMD), 26 as refractory anemia with excess blasts-1 (RAEB-1), 46 as refractory anemia with excess blasts-2 (RAEB-2), nine as unclassified myelodysplastic syndrome (MDS-U), and seven as 5q-syndrome. In the prognostic evaluation of MDSs, IPSS-R was a commonly used method. IPSS-R was based on these

TABLE 1 | Baseline characteristics of 376 cases in study.

Characteristic	*De novo* MDS	Pos ctl		NC	ICUS
		AML	MPN		
Number	200	63	25	26	62
Age, years	12–87	11–91	50–87	26–74	9–74
Median	60	61	71	55	62
Male/Female	115/85	32/19	14/11	13/13	32/30
WBC, $\times 10^9$/L	0.4–38.2	0.2–265.9	2.3–24.5	4.3–9.5	1.2–11.5
Median	3.6	8.1	6.7	6.7	5.6
Hb, g/L	27–168	38–147	65–187	123–146	34–132
Median	82	89	102	132	66
PLT, $\times 10^9$/L	2–531	3–267	34–863	102–278	13–258
Median	96	42	167	176	71
Follow-up, months	6–42	8–39	6–40	–	6–42
Median	28	26	27	–	27

MDS, myelodysplastic syndromes; AML, acute myeloid leukemia; MPN, myeloproliferative neoplasm; ICUS, idiopathic cytopenia of undetermined significance; WBC, white blood cell; Hb, hemoglobin; PLT, platelet; Pos ctl, positive control; NC, normal control.

characteristics (depth of cytopenias, splitting of marrow blasts <5%, and more precise cytogenetic subtypes). MDS patients were more precisely classified into all five IPSS-R categories, including Very low, Low, Intermediate, High, and Very high subgroups. Cytogenetic results accounted for an important proportion and could be divided into five categories, including Very good [–Y, del(11q)], Good [Normal, del(5q), del(12p), del(20q), double including del(5q)], Intermediate [del(7q), +8, +19, i(17q), any other single or double independent clones], Poor [−7, inv(3)/t(3q)/del(3q), double including −7/del(7q), complex: three abnormalities], and Very poor (complex: >3 abnormalities) subtypes (18). According to the IPSS-R standard, MDS patients in each subgroup were 10, 41, 54, 55, and 26, respectively; However, there were 14 cases not classified due to no cell growth available for MC analysis. The clinical features of these subgroups have been presented in **Supplementary Table 1**. The lower-risk group consisted of patients from the Very low, Low, and Intermediate categories of IPSS-R, and the higher-risk group was composed of patients from the High and Very high categories of IPSS-R. Patients were considered for clinical management driven by individual patient's clinical and biological characteristics and by physician preferences. Patients were managed according to the Chinese Expert Consensus on Diagnosis and Treatment of MDS (19). The goal of treatment for low-risk MDS patients was to improve the quality of life. The treatment was mainly supportive care, including blood transfusion, erythropoietin (EPO) and granulocyte colony-stimulating factor (G-CSF) administration, and removal of iron. Commonly used immunomodulation therapy drugs include thalidomide and lenalidomide. The target of MDS treatment in high-risk groups was to delay disease progression, prolong survival, and cure. The high-risk patients were treated with decitabine and/or chemotherapy. Hematopoietic stem cell transplantation was performed in eight of our patients.

All 376 recruited cases were subjected to SNP-A and MC studies on their BM samples. All samples were obtained at disease presentation.

This work was prospectively conducted in regard to specimen collection and clinical follow-up. OS was measured from day 0 to death from any cause (patients lost to follow-up were censored). PFS was defined as the time from day 0 to disease progression. This study was approved by the Ethics Committee of Tianjin Medical University General Hospital and Tianjin First Central Hospital. Patients and healthy controls gave their informed consent. The study was conducted in accordance with the Declaration of Helsinki.

Cytogenetic Analysis

Cytogenetic analysis of bone marrow aspirates was performed according to standard methods. The chromosomal preparations were G-banded using trypsin and Giemsa (GTG), and the karyotypes were described according to the International System for Human Cytogenetic Nomenclature (ISCN) (20).

Single-Nucleotide Polymorphism Array Analysis

SNP-A analysis was performed at Wuhan Kindstar Diagnostics Co./Kindstar Global gene (Beijing) Technology, Inc., P. R. China, by using the GeneChip Mapping 750K Assay Kit (CytoScan® 750K Assay Kit, Affymetrix, USA). Testing procedures were performed in strict accordance with the manufacturer's instructions and quality control standards, primarily including the steps of DNA extraction, enzyme digestion, connection, PCR, purification, fragmentation, labeling, hybridization, scanning, and data analysis. The detection instrument used was the GCS 3000Dx v.2 gene chip system, which is certified by the FDA/CE/CFDA, and the software used for data analysis was ChAS. The CytoScan 750K chip employed has more than 750,000 probes coated for the detection of genomic variance and covers 4,127 genes that include all the ISCA (International Standards for Cytogenomic Arrays) genes and 83% of the OMIM (Online Mendelian Inheritance in Man) disease-related genes. This chip can reliably detect copy number variations (CNVs), UPDs, and >10% of abnormal clones in mosaicism but is incapable of detecting balanced chromosome rearrangements and DNA point mutations. In the present study, three criteria were used to interpret a significant genomic aberration: First, the size of an identified aberration should be ≥400 Kb (for a gain), ≥400 Kb (for a loss), or ≥5 Mb (for a UPD) based on the manufacturer's recommendation and our own database. Second, the frequency of the identified aberration should be somewhat in concordance with the percentage of BM blasts in a patient, which could suggest that the aberration is likely acquired instead of constitutional in nature. Therefore, only aberrations in mosaic status (>10% of abnormal clones) were employed for further investigations. A threshold of 10% for mosaic identification was validated and provided by the manufacturer. Last, with regard to whether the aberration had been reported in association with respected disorders, related literature, and the Atlas of Genetics and Cytogenetics in Oncology and Hematology (http://atlasgeneticsoncology.org/

Anomalies/Anomliste.html) should be reviewed and checked to identify possible disease relationships.

Statistical Analysis

Categorical variables were compared using Fisher's exact test and the χ^2 test. Variance analysis was used to compare measurement data. Survival analysis was performed using the Kaplan–Meier method, and the Cox proportional hazard model was used for univariate analysis and multivariate analysis. All *P*-values are two-tailed, and $P < 0.05$ indicates statistical significance. Statistical analyses were performed with SPSS version 19.0.

RESULTS

Single-Nucleotide Polymorphism Array Analysis Led to a Higher Detection Rate of Chromosome Abnormalities

Our evaluation was performed on 376 cases that had been referred for identification of chromosome abnormalities by MC and SNP-A methods (**Supplementary Table 2**). MC allowed for the detection of 17 balanced rearrangements that were not detected by SNP-A. However, all the unbalanced chromosome aberrations identified by MC were also detected by SNP-A. In addition, SNP-A was able to detect many submicroscopic or cryptic chromosome abnormalities, which could not be detected by MC. The abnormality detection rate by SNP-A was 73.5, 72, and 69.8%, but by MC, it was 42, 48, and 36.5% in MDS, MPN, and AML patients, respectively. Comparing the two groups, the *P*-values were $P \leq 0.001$, $P = 0.148$, and $P \leq 0.001$, respectively. Notably, in our positive controls, the abnormal detection rates by both MC and SNP-A were higher in the MPN patients than in the AML patients likely due to the relatively small number of MPN patients enrolled in the study. Because our MPN and AML patients served as the positive controls, their detection results are only provided for assay validation purposes.

Importantly, in the 20 combined cases of MDS, MPN, and AML that had no informative MC findings (no cell growth available for MC analysis), 11 (55%) were found to be abnormal by SNP-A. In addition, with SNP-A analysis, the detection rates of submicroscopic or cryptic aberrations in the MDS, MPN, and AML patients with normal or no informative MC findings were 32.8, 30.8, and 30%, respectively. Furthermore, SNP-A-based aberrations in addition to the detection of MC in a patient were observed in 31% of the MDS, 50% of the MPN, and 30.4% of the AML patients. Notably, there were no abnormalities as detected by either MC or SNP-A in the normal controls.

Finally, even though all 62 ICUS patients were found to be normal by MC, 20 of them (32.2%) were identified as abnormal according to the SNP-A analysis.

Single-Nucleotide Polymorphism Array Analysis Revealed More Complex Chromosome Abnormalities

Using SNP-A, both CNVs and UPDs were observed in our MDS patients, with chromosome gains accounting for 42.0%, losses for 38.4%, and UPDs for 19.6%. The number of CNVs per patient ranged from 0 to 15, with a median number of 2.0 CNVs/patient. Notably, 88 of the 147 (59.9%) MDS patients with abnormal SNP-A detections showed 1–2 CNVs per patient, and 59 of the 147 (40.1%) showed ≥3 CNVs per patient. The SNP-A-detected abnormalities were found to involve essentially all 24 chromosomes, with chromosomes 1, 5, 7, 8, 9, 12, 17, 18, 19, 20, and 21 being affected relatively frequently. The detected chromosome aberrations by SNP-A mainly appeared as Gain 1q21, Loss 5q11, Loss 5q14, Loss5, Loss 7q11, Loss 7q22, Loss 7p21, Gain 8, Gain 9p13, Loss 9q21, UPD 9q21, Loss 12p11, Loss 12p13, Loss 17p11, Loss 17p13, Loss 18p11, Gain

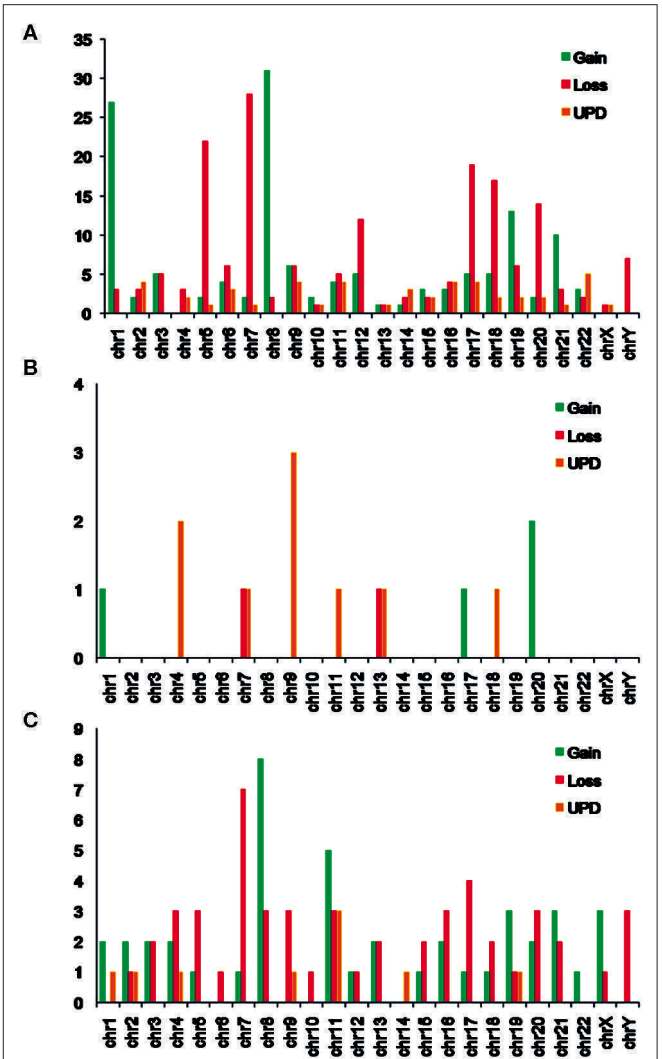


FIGURE 1 | (A) Number of myelodysplastic syndrome (MDS) patients with different types of single-nucleotide polymorphism array (SNP-A) abnormalities in each chromosome. **(B)** Number of myeloproliferative neoplasm (MPN) patients with different types of SNP-A abnormalities in each chromosome. **(C)** Number of acute myeloid leukemia (AML) patients with different types of SNP-A abnormalities in each chromosome.

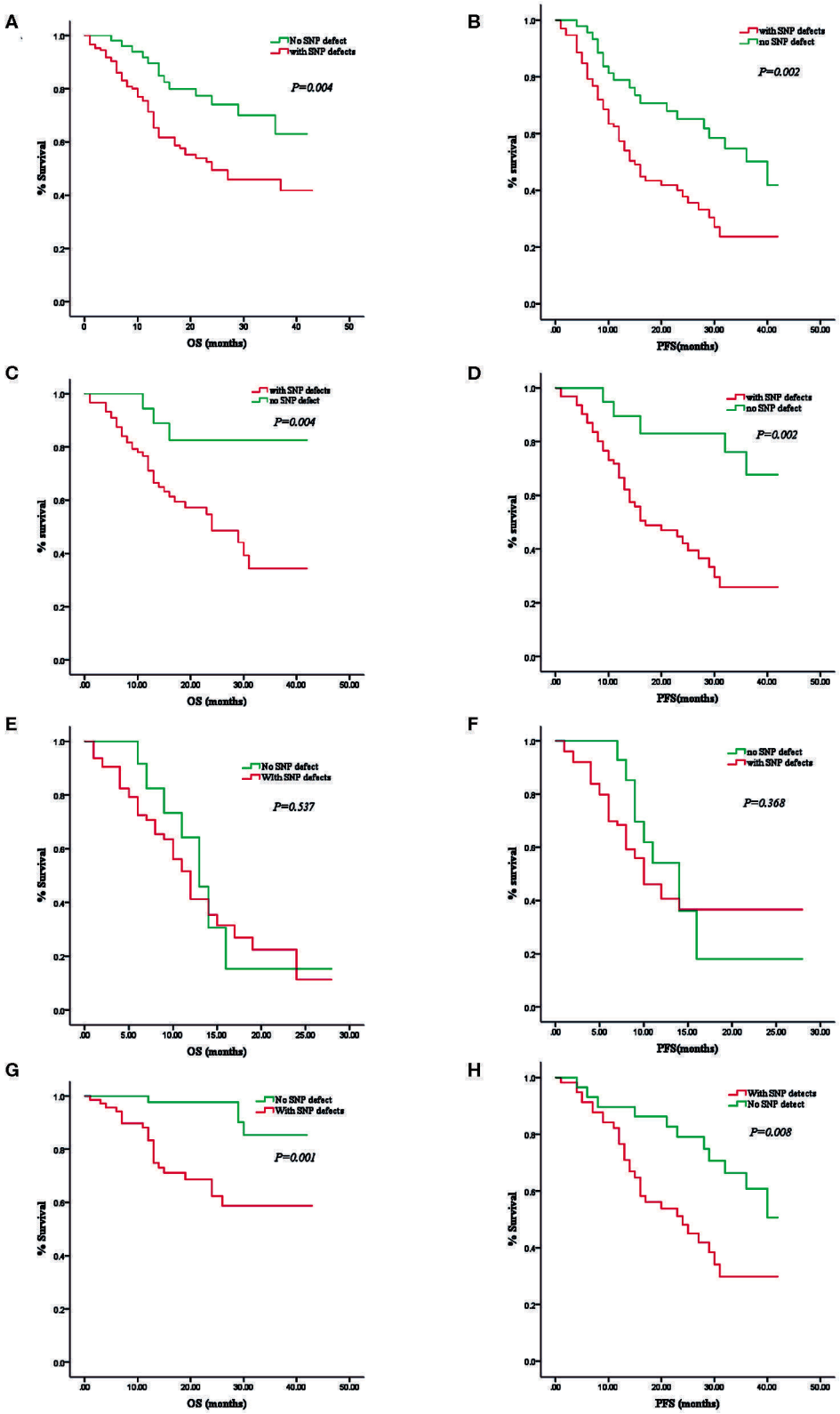


FIGURE 2 | Correlations between overall survival (OS)/progression-free survival (PFS) and single-nucleotide polymorphism array (SNP-A) detections in patients with myelodysplastic syndrome (MDS). Comparison of the MDS patients with and without SNP-A aberrations in OS **(A)** and PFS **(B)**. Comparison of the MDS patients with abnormal SNP-A detections and without such additional SNP-A aberrations in OS **(C)** and PFS **(D)** of the normal or good cytogenetic findings by metaphase cytogenetics (MC). Comparison of the MDS patients with and without SNP-A aberrations in OS **(E)** and PFS **(F)** of the high-risk group. Comparison of the MDS patients with and without SNP-A aberrations in OS **(G)** and PFS **(H)** of the low-risk group.

19p13, Loss 19p13, Loss 20q11, and Loss 20q12. Notably, UPDs were observed to involve chromosomes 2, 4, 6, 9, 11, 19, and 22 (**Figure 1A**). All these findings were largely consistent with previously reported observations (2, 3, 5, 9).

In our positive controls (MPN and AML patients), many chromosomal abnormalities were also observed by SNP-A. Notably, these abnormalities were identified as commonly involving chromosomes 4, 7, 9, 13, and 20 in the MPN patients and chromosomes 7, 8, 11, and 17 in the AML patients (**Figures 1B,C**).

Chromosomal Aberrations Detected by Single-Nucleotide Polymorphism Array Contributed to a Poor Prognosis in Patients With Myelodysplastic Syndromes

IPSS-R evaluation predicts overall survival and leukemia-free survival of patients with primary MDSs (18). There is no doubt that cytogenetics is one of the most valuable indicators in assessing MDS prognosis in the "gold standard" scoring system. In our study, except for seven patients lost to follow-up, the remaining 193 patients with MDS were followed up for 6–42 months with a median time of 28 months. The MDS patients with SNP-A-detected abnormalities had significantly lower OS (24 months vs. NR; $P = 0.004$) and PFS (15 vs. 40 months; $P = 0.002$) than those without SNP-A abnormalities (**Figures 2A,B**). In addition, we evaluated the prognostic value of SNP-A analysis in MDS patients with normal karyotypes or good IPSS-R karyotypes by MC. Of these patients, the prognosis of the patients with abnormal SNP-A detections was significantly worse in terms of OS and PFS (**Figures 2C,D**).

According to the IPSS-R standard, high-risk and very-high-risk MDS patients were classified as the high-risk group, and very-low-risk, low-risk, and intermediate-risk MDS patients were classified as the low-risk group. In our study, SNP-A analysis did not demonstrate an advantage in prognostic assessment for the high-risk group (**Figures 2E,F**). However, in the low-risk group, the patients with abnormal SNP-A detections had a significantly shorter survival time than patients without SNP-A aberrations (**Figures 2G,H**). Therefore, for MDS patients with a low-risk evaluation according to IPSS-R, SNP-A analysis seems to have a more significant impact on prognostic prediction.

Finally, in one patient, the number of SNP-A abnormalities, clinical features (including sex, age, blood counts, bone marrow blasts), and MC findings were also used to evaluate the prognosis of MDS patients by multivariable analysis (**Table 2**). The number of SNP-A abnormalities (≥3 per patient) was an independent predictor of poor prognosis in the patients with MDS [hazard ratio (HR) = 2.40, $P = 0.002$]. Our investigations provided valuable additional risk-stratification information to the standard IPSS-R scoring system.

TABLE 2 | Multivariable analysis of clinical data, MC findings, and number of SNP-A aberrations.

Factor	Hazard ratio (95% CI)	*P*
Age	1.73 (0.75–4.07)	0.002
Sex (male vs. female)	1.47 (1.01–1.69)	0.007
NEU ($\times 10^9$/L) (<0.8 vs. ≥0.8)	1.19 (0.81–2.92)	0.029
Hb (g/L) (<80 vs. 80–100 vs. ≥100)	1.52 (1.06–4.02)	0.016
Plt ($\times 10^9$/L) (<50 vs. 50–100 vs. ≥100)	1.06 (0.58–1.52)	0.030
BM blasts (%) (<5 vs. 5–10 vs. >10)	1.79 (1.04–3.47)	0.016
MC (very good, good, intermediate vs. poor, very poor)	2.22 (0.79–6.12)	0.008
Number of SNP-A aberrations (≥3 vs. <3)	2.40 (1.48–9.57)	0.002

NEU, neutrophil; Hb, hemoglobin; Plt, platelet; BM, bone marrow; MC, metaphase cytogenetics; SNP-A, single nucleotide polymorphism array.

Chromosomal Aberrations Detected by Single-Nucleotide Polymorphism Array Were Closely Associated With a High Risk of Transformation to Typical Myelodysplastic Syndrome in Patients With Idiopathic Cytopenia of Undetermined Significance

Patients with ICUS may progress to myeloid malignancies, but most patients often maintain a stable ICUS status for many years without progression. An ICUS patient once identified as having an abnormal karyotype that meets the MDS criteria would be rediagnosed with MDS. SNP-A can efficiently detect chromosomal abnormalities, which is important for assessing the evolution of the disease. In our study, 20 of the 62 ICUS patients were found to have chromosomal abnormalities by SNP-A technology. These abnormalities affected almost all chromosomes except chromosomes 2, 10, 11, 13, 16, and X (**Table 3**). These 20 ICUS patients with SNP-A aberrations were followed up for a median of 11 months (6–20 months). Notably, 17 of them (85%) transformed to typical MDS, and the remaining three (15%) transformed to aplastic anemia (AA) (**Table 3**). However, the other 42 ICUS patients without SNP-A abnormalities were also followed up for a median of 12 months (3–24 months), and none of them were converted to MDS. Therefore, chromosomal abnormalities detected by SNP-A were closely associated with a high risk of disease transformation in patients with ICUS.

DISCUSSION

The global profiling of DNA copy number changes in cancer cells through the use of microarray platforms is extremely attractive because it provides an unparalleled opportunity to uncover elusive genomic aberrations that are critical to tumorigenesis and progression. SNP-A technology allows for the capture of DNA copy number changes and SNP-based genotypes at sub base

TABLE 3 | Aberrations detected by SNP-A in 20 ICUS patients.

Patients	Aberrations	Diagnosis*	Time**
1	UPD (17q11.1-q11.2)	RAEB-1	8 months
2	Loss(20q), Gain(21q), UPD(14q)	MDS-U	12 months
3	Loss(5q21.1-qter), Loss(12p), Loss(17q)	RCMD	6 months
4	Loss(Y)	AA	13 months
5	UPD(19p)	AA	8 months
6	UPD(6p)	RCUD	10 months
7	Gain(8)	RCUD	12 months
8	UPD(14q)	RCMD	10 months
9	Gain(1q)	RCMD	15 months
10	Loss(20q)	RCUD	13 months
11	Loss(3p), Gain(18q), UPD(9p,12q)	RCMD	6 months
12	UPD(19q)	RCUD	14 months
13	Gain(1q), Loss(7q), UPD(15q,17q)	RCMD	8 months
14	Gain(8)	MDS-U	10 months
15	Loss(Y)	RCUD	14 months
16	UPD(4q)	RCMD	18 months
17	Gain(8)	AA	9 months
18	UPD(4q)	RCUD	17 months
19	Loss(4q,5q,11p,17), Gain(21q)	RAEB-1	7 months
20	UPD(5q)	RCUD	20 months

UPD, uniparental disomy; RAEB-1, refractory anemia with excess blasts-1; MDS-U, myelodysplastic syndrome, unclassified; RCMD, refractory cytopenia with multilineage dysplasia; RCUD, refractory cytopenia with unilineage dysplasia; AA, aplastic anemia.
**Diagnosis after transformation from ICUS.*
***Follow-up time from initial diagnosis to disease transformation.*

resolution, which helps detect small-scale genomic lesions and UPDs. A series of SNP-A-based studies have been performed on hematologic disorders, including acute lymphoblastic leukemia (21), MDS (22–25), myeloma (26), leukemias (27–29), and lymphomas (30).

From a technological point of view, our investigations have demonstrated that the detection of chromosomal abnormalities can be improved significantly by using the SNP-A technique for patients with MDS. From the following several aspects of data analyses, even somewhat confirmatory for previous findings in nature, we could still better appreciate the technical advantages of SNP-A over MC in detecting chromosomal aberrations. First, in our study, the abnormal detection rate by SNP-A for the patients with MDS and for the positive controls (MPN and AML patients) was higher than that obtained by MC. Second, SNP-A allowed for the detection of cryptic chromosomal lesions in the MDS patients and the positive controls with normal, abnormal, or even no informative MC findings, meaningfully demonstrating the technical reliability of SNP-A analysis. Third, SNP-A can detect chromosome deletions, gains, and UPDs. Acquired UPDs have been described in several malignancies (31–33), but due to the inability of MC to identify them, UPDs have remained largely elusive in many hematological disorders. Acquired segmental UPD is likely the result of mitotic recombination and appears to be a common event in MDS (24, 34, 35). In our study, acquired UPDs were observed in 19.6% of the MDS patients, with chromosomes 2, 4, 6, 9, 11, 19, and 22 being involved, which is largely consistent with previous reports. Finally, from a practical point of view, we would still recommend the combined application of MC and SNP-A for detection because MC can offset the inability of SNP-A to identify balanced chromosome rearrangements.

From a clinical point of view, our studies offered the following findings either not previously reported or less emphasized:

(1) Remarkably, in our study, 20 of the 62 ICUS patients had abnormal SNP-A detections, and 17 of these 20 patients progressed to typical MDSs with a progression time of 6–20 months and a median progression time of 11 months. Thus, abnormal SNP-A detections may predict the transformation to MDSs in advance for patients with ICUS, which would lead to disease monitoring and early intervention.

(2) It is likely that the presence of chromosome abnormalities as detected by SNP-A is responsible for the prediction of clinical phenotype and prognosis. A series of studies have shown that SNP-A detection is closely associated with prognosis (24–26). In this regard, our current study further strengthened the clinical value of SNP-A detection in prognostic assessment for patients with MDS. As a result, the patients with a normal SNP-A finding likely had a more favorable prognosis; SNP-A detection had an especially important value for prognostic assessment of the MDS patients in the low-risk group; the number of abnormalities ($\geq$3 per patient) was observed to be an independent predictor of poor prognosis. Therefore, our observations are of significant clinical value and provide additional information important for further risk-stratification assessment of patients with MDSs. Based on our findings and those of previous reports, it is now evident that a combination of MC and SNP-A methods would provide a more precise assessment of the prognosis of patients with MDSs. Recently, a series of studies (2, 6, 22, 36) showed that total genomic alterations detected by SNP-A were predictive of overall survival in a cohort of patients with MDSs or other related hematological disorders who received demethylation-based treatment, which certainly deserves further investigation.

A better understanding of the strength and weakness of each technique in a clinical setting is of extreme importance. SNP-A can detect loss of heterozygosity and serve as a useful complement to MC by capturing additional submicroscopic or cryptic chromosome gains or deletions. However, SNP-A can only detect chromosomal or chromosome-fragment-size aberrations but cannot detect single gene-based mutations. Recently, Choi et al. (37) used a more sensitive SNP-A approach (Affymetrix CytoScan HD) to investigate submicroscopic or cryptic chromosome aberrations in MDS patients. This CytoScan HD platform had ~2.7 million coated probes (much more than that of the CytoScan 750K chip employed in our study) and was able to detect gains or losses of more than 35 markers within or including a known clinically significant cancer-related gene. Thus, in the study by Choi et al., they could identify much smaller cryptic abnormalities, such as KMT2A partial tandem duplication and deletion involving the TET2 gene, that are often smaller than 100 kb in size. Certainly, the

CytoScan 750K-based SNP-A platform adopted in our study cannot reach such a greater sensitivity in detection. Based on the detection of chromosome-fragment sized aberrations (often >400 kb in size), our study provided several findings either not previously reported or less emphasized as described above and should be considered valuable information complementary to Choi's findings. Next-generation sequencing (NGS) focuses more on gene mutation analysis. Mutant genes can be detected in more than 80% of MDS patients, and most mutations are not specific and usually have uncertain significance (38). Although NGS makes it increasingly easy to detect fusions and mutations, not all cytogenetic abnormalities can be detected by NGS. Therefore, if feasible, these techniques should be combined to contribute to the study of genomic aberrations for better and more precise management of patients with MDS (39–42).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was approved by the Ethic committee of Tianjin Medical University General Hospital and Tianjin First Central Hospital. Patients and healthy controls gave their informed consent. The study was conducted in accordance with the Declaration of Helsinki.

AUTHOR'S NOTE

Presented in abstract form at the 59th annual meeting of the American Society of Hematology, Atlanta, GA, December 9, 2017. TITLE: Chromosome aberrations detected by SNP array technique indicating a high risk of MDS transformation in patients with ICUS and poor prognosis of patients with MDS.

AUTHOR CONTRIBUTIONS

MZ, ZC, and ZS designed the study. XX and XH collected and analyzed the data and wrote the manuscript. QL, WZ, HZ, WY, YL, LG, HL, LL, HW, and RF provided clinical data. MZ, ZC, and ZS reviewed the manuscript and contributed to the final draft. All authors contributed to the article and approved the submitted version.

FUNDING

We thank Affymetrix Inc. (USA) for financial support. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. We thank Wuhan Kindstar Diagnostics Co./Kindstar Global gene (Beijing) Technology, Inc., for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00962/full#supplementary-material

REFERENCES

1. Killick SB, Carter C, Culligan D, Dalley C, Das-Gupta E, Drummond M, et al. Guidelines for the diagnosis and management of adult myelodysplastic syndromes. *Br J Haematol.* (2014) 164:503–25. doi: 10.1111/bjh.12694
2. Arenillas L, Mallo M, Ramos F, Guinta K, Barragan E, Lumbreras E, et al. Single nucleotide polymorphism array karyotyping: a diagnostic and prognostic tool in myelodysplastic syndromes with unsuccessful conventional cytogenetic testing. *Genes Chromosomes Cancer.* (2013) 52:1167–77. doi: 10.1002/gcc.22112
3. Tiu RV, Gondek LP, O'Keefe CL, Elson P, Huh J, Mohamedali A, et al. Prognostic impact of SNP array karyotyping in myelodysplastic syndromes and related myeloid malignancies. *Blood.* (2011) 117:4552–60. doi: 10.1182/blood-2010-07-295857
4. Sole F, Luno E, Sanzo C, Espinet B, Sanz GF, Cervera J, et al. Identification of novel cytogenetic markers with prognostic significance in a series of 968 patients with primary myelodysplastic syndromes. *Haematologica.* (2005) 90:1168–78.
5. Cluzeau T, Moreilhon C, Mounier N, Karsenti JM, Gastaud L, Garnier G, et al. Total genomic alteration as measured by SNP-array-based molecular karyotyping is predictive of overall survival in a cohort of MDS or AML patients treated with azacitidine. *Blood Cancer J.* (2013) 3:e155. doi: 10.1038/bcj.2013.52
6. Ganster C, Shirneshan K, Salinas-Riester G, Braulke F, Schanz J, Platzbecker U, et al. Influence of total genomic alteration and chromosomal fragmentation on response to a combination of azacitidine and lenalidomide in a cohort of patients with very high risk MDS. *Leuk Res.* (2015) 39:1079–87. doi: 10.1016/j.leukres.2015.06.011
7. Pozdnyakova O, Miron PM, Tang G, Walter O, Raza A, Woda B, et al. Cytogenetic abnormalities in a series of 1,029 patients with primary myelodysplastic syndromes: a report from the US with a focus on some undefined single chromosomal abnormalities. *Cancer.* (2008) 113:3331–40. doi: 10.1002/cncr.23977
8. Haase D, Germing U, Schanz J, Pfeilstocker M, Nosslinger T, Hildebrandt B, et al. New insights into the prognostic impact of the karyotype in MDS and correlation with subtypes: evidence from a core dataset of 2124 patients. *Blood.* (2007) 110:4385–95. doi: 10.1182/blood-2007-03-082404
9. da Silva FB, Machado-Neto JA, Bertini V, Velloso E, Ratis CA, Calado RT, et al. Single-nucleotide polymorphism array (SNP-A) improves the identification of chromosomal abnormalities by metaphase cytogenetics in myelodysplastic syndrome. *J Clin Pathol.* (2017) 70:435–42. doi: 10.1136/jclinpath-2016-204023
10. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood.* (2009) 114:937–51. doi: 10.1182/blood-2009-03-209262
11. Teh MT, Blaydon D, Chaplin T, Foot NJ, Skoulakis S, Raghavan M, et al. Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event. *Cancer Res.* (2005) 65:8597–603. doi: 10.1158/0008-5472.CAN-05-0842
12. Gaasenbeek M, Howarth K, Rowan AJ, Gorman PA, Jones A, Chaplin T, et al. Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers. *Cancer Res.* (2006) 66:3471–9. doi: 10.1158/0008-5472.CAN-05-3285

13. Mohamedali A, Gaken J, Twine NA, Ingram W, Westwood N, Lea NC, et al. Prevalence and prognostic significance of allelic imbalance by single-nucleotide polymorphism analysis in low-risk myelodysplastic syndromes. *Blood.* (2007) 110:3365–73. doi: 10.1182/blood-2007-03-079673
14. Maciejewski JP, Tiu RV, O'Keefe C. Application of array-based whole genome scanning technologies as a cytogenetic tool in haematological malignancies. *Br J Haematol.* (2009) 146:479–88. doi: 10.1111/j.1365-2141.2009.07757.x
15. Valent P, Horny HP. Minimal diagnostic criteria for myelodysplastic syndromes and separation from ICUS and IDUS: update and open questions. *Eur J Clin Invest.* (2009) 39:548–53. doi: 10.1111/j.1365-2362.2009.02151.x
16. Valent P, Horny HP, Bennett JM, Fonatsch C, Germing U, Greenberg P, et al. Definitions and standards in the diagnosis and treatment of the myelodysplastic syndromes: consensus statements and report from a working conference. *Leuk Res.* (2007) 31:727–36. doi: 10.1016/j.leukres.2006.11.009
17. Cannella L, Breccia M, Latagliata R, Frustaci A, Alimena G. Clinical and prognostic features of patients with myelodysplastic/myeloproliferative syndrome categorized as unclassified (MDS/MPD-U) by WHO classification. *Leuk Res.* (2008) 32:514–6. doi: 10.1016/j.leukres.2007.07.004
18. Greenberg PL, Tuechler H, Schanz J, Sanz G, Garcia-Manero G, Sole F, et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood.* (2012) 120:2454–65. doi: 10.1182/blood-2012-03-420489
19. Depei. W, Changgeng. R, Xiaojun. H. [Expert consensus on diagnosis and treatment of myelodysplastic syndrome (2014)]. *Zhonghua Xue Ye Xue Za Zhi.* (2014) 35:1042–8. doi: 10.3760/cma.j.issn.0253-2727.2014.11.023
20. Schreck R, Distèche C, Adler D. ISCN standard idiograms. *Curr Protoc Hum Genet.* (2001) Appendix 4:Appendix 4B. doi: 10.1002/0471142905.hga04bs18
21. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature.* (2007) 446:758–64. doi: 10.1038/nature05690
22. Yi JH, Huh J, Kim HJ, Kim SH, Kim SH, Kim KH, et al. Genome-wide single-nucleotide polymorphism array-based karyotyping in myelodysplastic syndrome and chronic myelomonocytic leukemia and its impact on treatment outcomes following decitabine treatment. *Ann Hematol.* (2013) 92:459–69. doi: 10.1007/s00277-012-1635-7
23. Heinrichs S, Li C, Look AT. SNP array analysis in hematologic malignancies: avoiding false discoveries. *Blood.* (2010) 115:4157–61. doi: 10.1182/blood-2009-11-203182
24. Svobodova K, Zemanova Z, Lhotska H, Novakova M, Podskalska L, Belickova M, et al. Copy number neutral loss of heterozygosity at 17p and homozygous mutations of TP53 are associated with complex chromosomal aberrations in patients newly diagnosed with myelodysplastic syndromes. *Leuk Res.* (2016) 42:7–12. doi: 10.1016/j.leukres.2016.01.009
25. Evans AG, Ahmad A, Burack WR, Iqbal MA. Combined comparative genomic hybridization and single-nucleotide polymorphism array detects cryptic chromosomal lesions in both myelodysplastic syndromes and cytopenias of undetermined significance. *Mod Pathol.* (2016) 29:1183–99. doi: 10.1038/modpathol.2016.104
26. Walker BA, Leone PE, Jenner MW, Li C, Gonzalez D, Johnson DC, et al. Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. *Blood.* (2006) 108:1733–43. doi: 10.1182/blood-2006-02-005496
27. Irving JA, Bloodworth L, Bown NP, Case MC, Hogarth LA, Hall AG. Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. *Cancer Res.* (2005) 65:3053–8. doi: 10.1158/0008-5472.CAN-04-2604
28. Gorletta TA, Gasparini P, D'Elios MM, Trubia M, Pelicci PG, Di Fiore PP. Frequent loss of heterozygosity without loss of genetic material in acute myeloid leukemia with a normal karyotype. *Genes Chromosomes Cancer.* (2005) 44:334–7. doi: 10.1002/gcc.20234
29. Fitzgibbon J, Smith LL, Raghavan M, Smith ML, Debernardi S, Skoulakis S, et al. Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias. *Cancer Res.* (2005) 65:9152–4. doi: 10.1158/0008-5472.CAN-05-2017
30. Nielaender I, Martin-Subero JI, Wagner F, Martinez-Climent JA, Siebert R. Partial uniparental disomy: a recurrent genetic mechanism alternative to chromosomal deletion in malignant lymphoma. *Leukemia.* (2006) 20:904–5. doi: 10.1038/sj.leu.2404173
31. Pei J, Kruger WD, Testa JR. High-resolution analysis of 9p loss in human cancer cells using single nucleotide polymorphism-based mapping arrays. *Cancer Genet Cytogenet.* (2006) 170:65–8. doi: 10.1016/j.cancergencyto.2006.05.002
32. Raghavan M, Lillington DM, Skoulakis S, Debernardi S, Chaplin T, Foot NJ, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res.* (2005) 65:375–8.
33. Andersen CL, Wiuf C, Kruhoffer M, Korsgaard M, Laurberg S, Orntoft TF. Frequent occurrence of uniparental disomy in colorectal cancer. *Carcinogenesis.* (2007) 28:38–48. doi: 10.1093/carcin/bgl086
34. O'Keefe C, McDevitt MA, Maciejewski JP. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood.* (2010) 115:2731–9. doi: 10.1182/blood-2009-10-201848
35. Gondek LP, Tiu R, O'Keefe CL, Sekeres MA, Theil KS, Maciejewski JP. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. *Blood.* (2008) 111:1534–42. doi: 10.1182/blood-2007-05-092304
36. Koh KN, Lee JO, Seo EJ, Lee SW, Suh JK, Im HJ, et al. Clinical significance of previously cryptic copy number alterations and loss of heterozygosity in pediatric acute myeloid leukemia and myelodysplastic syndrome determined using combined array comparative genomic hybridization plus single-nucleotide polymorphism microarray analyses. *J Korean Med Sci.* (2014) 29:926–33. doi: 10.3346/jkms.2014.29.7.926
37. Choi SM, Van Norman SB, Bixby DL, Shao L. Cytogenomic array detects a subset of myelodysplastic syndrome with increased risk that is invisible to conventional karyotype. *Genes Chromosomes Cancer.* (2019) 58:756–74. doi: 10.1002/gcc.22783
38. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood.* (2013) 122:3616–27; quiz 36399. doi: 10.1182/blood-2013-08-518886
39. Bono E, McLornan D, Travaglino E, Gandhi S, Galli A, Khan AA, et al. Clinical, histopathological and molecular characterization of hypoplastic myelodysplastic syndrome. *Leukemia.* (2019) 33:2495–505. doi: 10.1038/s41375-019-0457-1
40. Tawana K, Drazer MW, Churpek JE. Universal genetic testing for inherited susceptibility in children and adults with myelodysplastic syndrome and acute myeloid leukemia: are we there yet? *Leukemia.* (2018) 32:1482–92. doi: 10.1038/s41375-018-0051-y
41. Gangat N, Mudireddy M, Lasho TL, Finke CM, Nicolosi M, Szuber N, et al. Mutations and prognosis in myelodysplastic syndromes: karyotype-adjusted analysis of targeted sequencing in 300 consecutive cases and development of a genetic risk model. *Am J Hematol.* (2018) 93:691–7. doi: 10.1002/ajh.25064
42. Kennedy JA, Ebert BL. Clinical implications of genetic mutations in myelodysplastic syndrome. *J Clin Oncol.* (2017) 35:968–74. doi: 10.1200/JCO.2016.71.0806

Conflict of Interest: WY and ZC were employed by Wuhan Kindstar Diagnostics Co./Kindstar Global gene (Beijing) Technology, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fgene.2020.00771

VisTCR: An Interactive Software for T Cell Repertoire Sequencing Data Analysis

*Qingshan Ni[1,2†], Jianyang Zhang[1,2†], Zihan Zheng[3†], Gang Chen[1,2], Laura Christian[4], Juha Grönholm[5], Haili Yu[1,2], Daxue Zhou[1,2], Yuan Zhuang[4], Qi-Jing Li[4] and Ying Wan[1,2]**

[1] Biomedical Analysis Center, Army Medical University, Chongqing, China, [2] Chongqing Key Laboratory of Cytomics, Chongqing, China, [3] Biowavelet Ltd., Chongqing, China, [4] Department of Immunology, Duke University Medical Center, Durham, NC, United States, [5] Molecular Development of the Immune System Section, NIAID Clinical Genomics Program, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States

Recent progress in high throughput sequencing technologies has provided an opportunity to probe T cell receptor (TCR) repertoire, bringing about an explosion of TCR sequencing data and analysis tools. For easier and more heuristic analysis TCR sequencing data, we developed a client-based HTML program (VisTCR). It has a data storage module and a data analysis module that integrate multiple cutting-edge analysis algorithms in a hierarchical fashion. Researchers can group and re-group samples for different analysis purposes by customized "Experiment Design File." Moreover, the VisTCR provides a user-friendly interactive interface, by all the TCR analysis methods and visualization results can be accessed and saved as tables or graphs in the process of analysis. The source code is freely available at https://github.com/qingshanni/VisTCR.

Keywords: T cell sequencing, analysis tool, data analysis, Graphic user interface, T cell repertoire

OPEN ACCESS

Edited by:
Longxiang Xie,
Henan University, China

Reviewed by:
Chuanlong Cui,
Rutgers Biomedical and Health Sciences, United States
Chunlong Zhang,
Harbin Medical University, China

***Correspondence:**
Ying Wan
wanying516@foxmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 October 2019*
Accepted: *29 June 2020*
Published: *21 July 2020*

Citation:
Ni Q, Zhang J, Zheng Z, Chen G, Christian L, Grönholm J, Yu H, Zhou D, Zhuang Y, Li Q-J and Wan Y (2020) VisTCR: An Interactive Software for T Cell Repertoire Sequencing Data Analysis. Front. Genet. 11:771. doi: 10.3389/fgene.2020.00771

INTRODUCTION

Breakthroughs made in the development of antibody-based treatments for autoimmune diseases and tumor immunotherapy in recent have fueled an as-yet unmet need for feasible personal immune monitoring platforms to evaluate adaptive immune response (Han et al., 2015). T cells are one of the most critical players of adaptive immunity, with diverse functions including cell killing, providing B cell help (and consequently boost specific antibody production), and cytokine secretion. By capturing the identity and relative size of T cell clones, T cell receptor (TCR)-Seq offers an opportunity to observe changes in the composition of the adaptive immune system at homeostasis or during pathogenic responses (Aris et al., 2018; Fahl et al., 2018; Jiang et al., 2018). Sorting and clonotyping of purified T cell populations, such as Tregs, has yielded insight into pathogenic populations and phenotypic changes in autoimmunity, while the clarification of the clonal dynamics of tumor-infiltrating $CD8^+$ T cells responsive to tumor neoantigens is under intensive study due to their positive association with enhanced prognosis. This additional dimension of immune monitoring thus extends our understanding of adaptive immunity, and has the potential to inform treatment decisions.

Facilitated in part by the decreasing cost of next-generation sequencing, T cell repertoire sequencing (TCR-Seq) data has been rapidly generated in recent years (Robins, 2013; Six et al., 2013; Newell and Davis, 2014; Hou et al., 2016). Many tools have also been developed for T cell sequencing data analysis. Some of these focus on sequence assembly, assignment to genomic V, D and J genes,

extraction of CDR3 regions and error correction, such as IgBlast (Ye et al., 2013), TCRKlass (Yang et al., 2014), Decombinator (Thomas et al., 2013), IMSEQ (Kuchenbecker et al., 2015), MiTCR (Bolotin et al., 2013), and MiXCR (Bolotin et al., 2015). Others provide global evaluation methods on the TCR sequencing data, such as ARResT/Interrogate (Bystry et al., 2017), ImmunExplorer (Schaller et al., 2015), VDJtools (Gardner et al., 2015), VDJviz (Bagaev et al., 2016), Vidjil (Duez et al., 2016), and tcR (Nazarov et al., 2015), providing different methods to gain biological and clinical understanding by diversity measurements, clonotype distribution, similarity analysis, etc. Many of these tools also offer different types of visualizations for a given analysis that emphasize distinct interpretations. For instance, VDJviz can generate individual-sample circus plots for VJ usage, while tcR offers radar plots to emphasize divergence in VJ segments across samples. Other features, such as clonotype clustering in VDJil, may be more rarely provided by an individual tool.

However, these initial clonotype extraction and final visualization tools tend to be separated, and not all of these tools are readily intercompatible. As such, performing a more complete analysis of TCR repertoires would require a user to piece several of these tools together in order to generate comprehensive visualizations. Furthermore, most of the current tools are primarily operated by a command line interface, and data interpretation from such interfaces may be challenging for some wet lab immunological researchers, who may require extensive assistance from computational bioinformaticians to generate these analysis. The nuances between, and functional impact of applying, different clonotype extraction methods in terms of downstream interpretation may also be confusing. To overcome this barrier, we have developed the VisTCR (Visual TCRSeq) software, an interactive platform with a graphical user interface (GUI) for simplified management and analysis of TCR sequencing data. Starting from raw sequencing data, VisTCR can be used to directly perform clonotype extraction and downstream analyses within a single data management framework. VisTCR leverages three of the most commonly used extraction methods to allow users to more easily explore their data, and investigate the differences that may result from applying distinct analysis pipelines across a broad range of downstream visualizations.

DESIGN AND IMPLEMENTATION

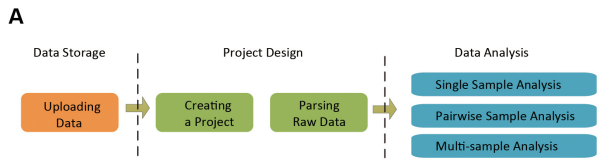
The design of VisTCR emphasizes a friendly, GUI and intuitive analysis workflow. The major features of the software include:

1. Independent modules for data management and analysis. In the Data Storage Module, raw data are uploaded and grouped in each sequencing experiments (**Figure 1B** and **Supplementary Video S1**). In the Data Analysis Module, the raw data can be selected and re-organized to perform various analyses and generate figures (**Figure 1C** and **Supplementary Video S2**).
2. Freedom to group samples for individualized analysis. An "Experiment Design File" is introduced in VisTCR that contains a combination of multiple variables for an analysis task, which allows users to de-construct their experiment data into a complex analysis design. Furthermore, in the data analysis process, individual variables or any combination of variables can be selected to group and re-group samples for comparison and analysis of T-cell sequencing data (**Supplementary Files S1, S2** and **Supplementary Video S2**).
3. Integration of multiple cutting-edge analysis algorithms in a hierarchical fashion. These data analysis methods in VisTCR are organized in hierarchical fashion and are divided into three categories: Single sample analysis, Pairwise samples analysis, and Multi-samples analysis. Each category is further subdivided to generate comprehensive repertoire analysis that includes visualizing clonotype distribution, similarity analysis and diversity analysis, and tracking individual clones across samples, etc. (**Figure 1A** and **Supplementary Table S1** and **Supplementary Video S2**).
4. User-friendly interactive interface and visualization of data. VisTCR provides a point and click interface for all of the TCR analysis methods. The analytical results are transformed into interactive data visualization with a representation-transparent approach (Bostock et al., 2011). These results can be downloaded as tables or graphs during each stage in the analysis workflow.

The workflow of VisTCR is composed of three steps (**Figure 1A**): (1) Uploading the sequencing data files into Data Storage Module, (2) Creating an analysis task in the Data Analysis Module, and (3) Performing analysis in Data Analysis Module. VisTCR use standard fastq format file as input, which is the most widely used format in sequence analysis. The raw TCR sequencing data files are uploaded, stored and organized in the "Experiment" tab of Data Storage Module (**Figure 1B** and **Supplementary Video S1**). A quality control tool (FastQC)[1] has been integrated to Data Storage Module for assessment of sequencing quality (**Supplementary Video S1**). In Data Analysis Module, an "Experiment Design File" is created firstly with a list of samples and variables to import the raw data from Data Storage Module into analysis workflow (**Supplementary Files S1, S2** and **Supplementary Video S2**). The raw TCR sequencing data can be parsed with several decoding methods [Decombinator (Thomas et al., 2013), MiTCR (Bolotin et al., 2013), and MiXCR (Bolotin et al., 2015)] as options (**Supplementary Figure S1**).

The analysis methods are categorized into three groups: Single sample analysis, Pairwise samples analysis, and Multi-samples analysis. In Single sample analysis, the TCRBV and/or TCRBJ usage, CDR3 spectratype and Clonotype distributions of selected samples can be analyzed. In Pairwise samples analysis, the shared clonotypes between two selected samples are shown in a plot with frequency of nucleotide or amino acid (nt/aa) sequences in Overlapping clonotype analysis. Moreover, the degeneracy of the shared T cell clonotypes is evaluated with Convergent Analysis, in which the number of unique CDR3 nucleotide sequences that are translated into same CDR3 amino acid sequence is calculated (Venturi et al., 2008). The Multi-sample analysis is classified into three categories: descriptive statistics, similarity analysis

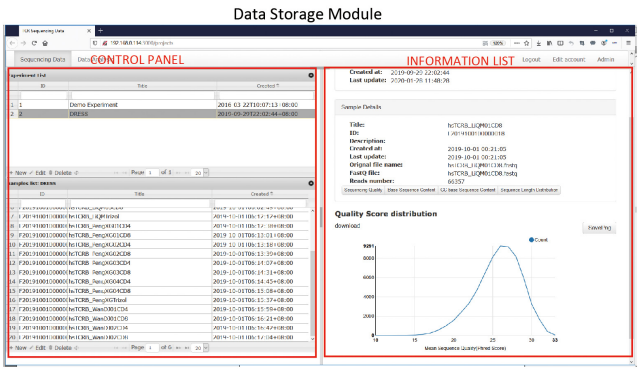
[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Comparison of web based software for T cell repertoire analysis

Analysis Items	VisTCR	VDJviz	Vidjil
TRBV Usage	●	●	●
TRBJ Usage	●	●	●
TRBV and TRBJ Usage	●	●	●
CDR3 Length distribution	●	●	●
Clonotype Frequency plots	●	○	●
Convergent analysis	●	○	○
search clonotypes	○	●	○
Rarefaction analysis	○	●	○
Overlapping Clonotype Frequency plots	●	○	○
Un-overlapping Clonotype Frequency plots	●	○	○
Most abundant clonotypes	●	○	○
Clonal space homeostasis	●	●	○
Clonotype tracking	●	●	●
Overlap analysis	●	○	○
Similarity in groups	●	○	○
Samples similarity matrix	●	○	○
Cluster analysis	●	○	●
Pairwise similarity analysis	●	○	○
Bio-diversity index	●	○	●
Significant test	●	○	○
Pairwise diversity analysis	●	○	○

● Included ○ Not included



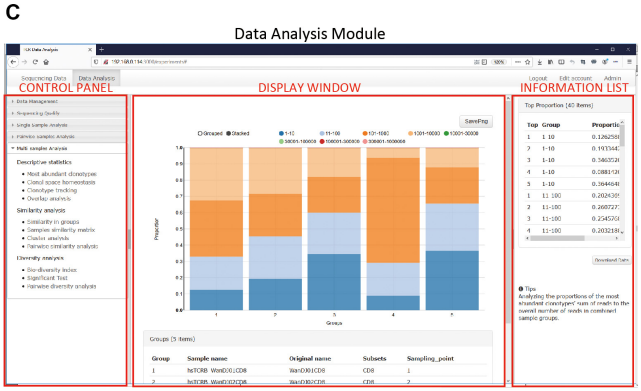


FIGURE 1 | Overview of the VisTCR software. **(A)** Basic workflow of the VisTCR software and TCR sequencing data analysis methods in VisTCR software. The workflow of the VisTCR software includes three steps (1) Uploading TCR sequencing data files in the Data Storage Module, (2) Creating an analysis project in the Data Analysis Module. (3) Performing analysis and obtaining results in the Data Analysis Module. **(B)** The GUI of the Data Storage Module. **(C)** The GUI of the Data Analysis Module.

and diversity analysis. The description statistics contain Most Abundant Clonotypes, Clonal Space Homeostasis, Clonotype Tracking, and Overlap Analysis. The similarity analysis and diversity analysis provide statistical methods to quantify the differences of grouped datasets by using a variety of similarity and diversity estimation methods (**Supplementary Table S1**). A list of the analyses that are possible in VisTCR with respect to two other commonly used tools featuring GUIs is also included for ease of comparison (**Figure 1A**). Notably, VisTCR enables a number of unique analyses for sequence convergence and clonotype overlap that are not available in the other tools.

The software is a client-based HTML program that has an intuitive user interface which is written in ROR (Ruby on Rails) (Bachle and Kirchberg, 2007), and Data-driven documents Javascript library (D3.js) (Bostock et al., 2011). The calculation is implemented using R language, which is integrated with ROR using Rserve[2].

RESULTS

To demonstrate the usage of VisTCR in T-cell repertoire analysis, a data set from a previously published paper was re-analyzed (Niu et al., 2015). As part of the original study to longitudinally characterize the $CD4^+/CD8^+$ T-cell repertoires in drug reaction with eosinophilia and systemic symptoms (DRESS) from diagnosis to clinical remission, $CD4^+$ and $CD8^+$ T-cells from peripheral blood of DRESS patients were isolated at 10-day intervals, and sequenced CDR3-regions of the TCRB chain on Ion Torrent PGM platform (Life Technologies, Carlsbad, CA, United States). This data set includes 66 samples from eight DRESS patient and 28 samples from healthy donors (Niu et al., 2015). All samples were uploaded into the data management module of VisTCR (**Supplementary Video S1**). Two experiment design files (**Supplementary Files S1, S2**) were edited to re-organize the data set. After uploading the experiment design files in the analysis module, two analysis tasks were created to demonstrate the cutting-edge analysis functions of VisTCR (**Supplementary Video S2**). One analysis task grouped the five timepoint TCR sequencing data from WDJ patient (**Supplementary File S1**). Another grouped the TCR sequencing data from the eight healthy donors together with samples taken at the first time pointfrom eight DRESS patients (**Supplementary File S2**). MiXCR with default parameters was used to extract CDR3 regions from raw sequences and perform error correction.

[2]http://www.rforge.net/Rserve/

Single Sample Analysis

The Single Sample Analysis in VisTCR was provided to browse the fundamental characters of TCR sequencing data to uncover clues for further analysis of each given sample. For instance, significant differences between the first and fifth timepoint data for the samples from patient WDJ (an obscured patient ID) could be found in terms of TRBV/J segment usage, CDR3 length distribution, and clonotype distribution could be observed from this analysis (**Supplementary Video S3** and **Figure 2**). The increase usage of TRBV27, TRBV13, TRBV18 and decreased usage of TRBV5-8, TRBV19 were discovered in the TRBV usages of the two timepoint data (**Figures 2A,B**). The peak of CDR3 length was 45 bp at the first timepoint and 42 bp by the fifth timepoint (**Figures 2C,D**). The highest frequency of TCR clonotype reached 10% in fifth timepoint, but had only reached 1.8% in first timepoint (**Figures 2E,F**). These resulting visualizations are thus consistent with the original conclusion that a portion of the CD8 + T cells were rapidly expanding in DRESS patients.

Pairwise Sample Analysis

To inspect the change of the repertoire of $CD8^+$ T cells in the development of DRESS, the first and fifth timepoint TCR sequencing data of WDJ patient were selected to analyze the distribution of overlapped and un-overlapped clonotype in the

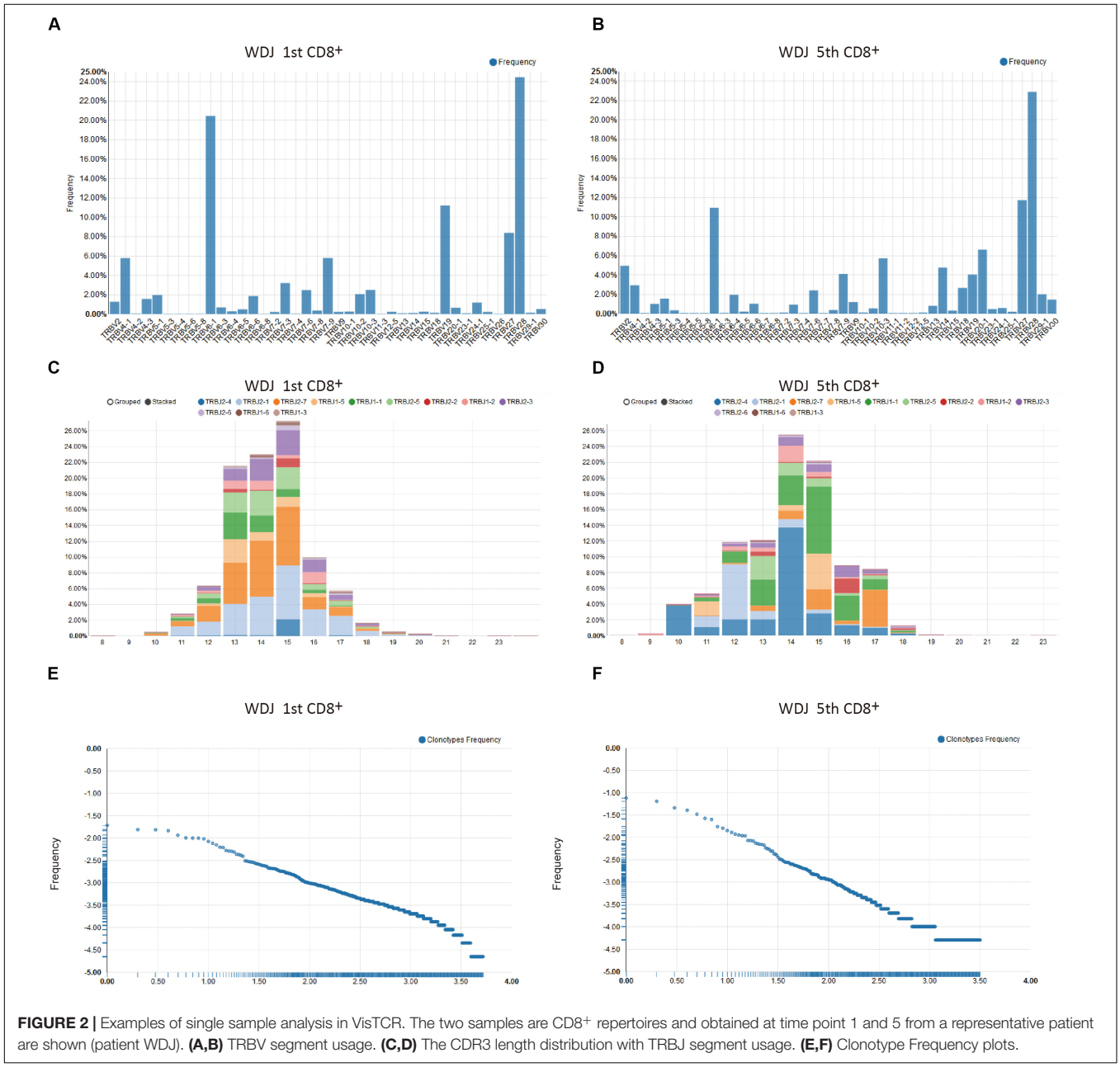


FIGURE 2 | Examples of single sample analysis in VisTCR. The two samples are $CD8^+$ repertoires and obtained at time point 1 and 5 from a representative patient are shown (patient WDJ). **(A,B)** TRBV segment usage. **(C,D)** The CDR3 length distribution with TRBJ segment usage. **(E,F)** Clonotype Frequency plots.

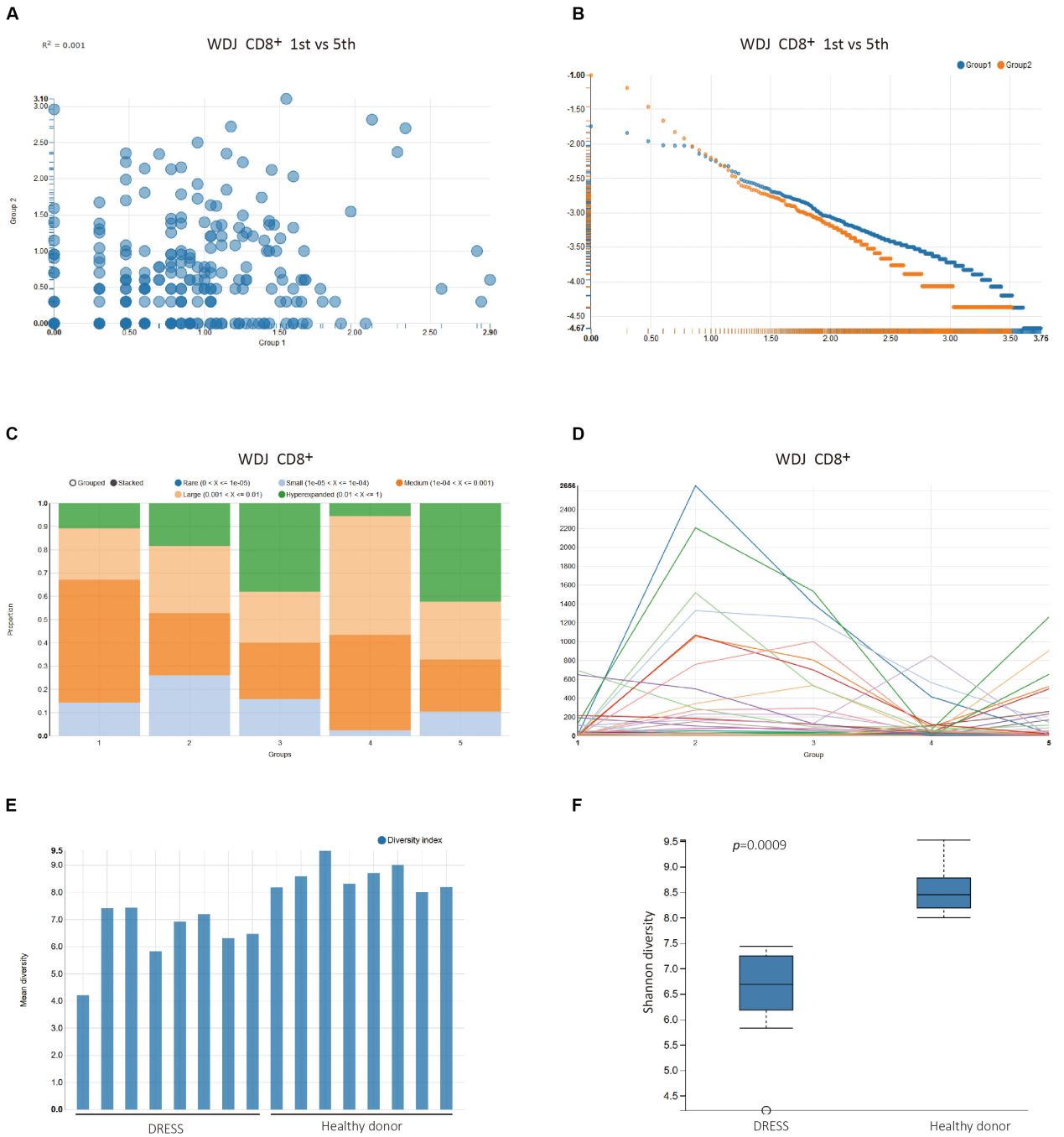


FIGURE 3 | Multi-time point sample analysis of patient WDJ. **(A)** Overlapping Clonotype Frequency plots between time point 1 (Group1) and time point 5 (Group 2). **(B)** Un-overlapping Clonotype Frequency plots of time point 1 (Group1) and time point 5 (Group 2). **(C)** Clonal space homeostasis visualization to help identify potential clonal expansion. The proportional distribution of the fourth timepoint TCR clonotypes differed from the third timepoint ($p < 0.0001$, Chi-square test) and fifth timepoint ($p < 0.0001$, Chi-square test). **(D)** Clonal tracking mapping the dominance of a given clone across all samples. Each line corresponds to a unique TCRB clonotype. As a general trend, it can be seen that a number of clones undergo clear expansion at the earlier timepoint (time2) before subsequently contracting (time4), a behavior consistent with memory T cell formation following the end of antigen exposure. **(E)** Bar plot of Shannon diversity index. Two groups, DRESS patients and healthy donors, of repertoires are selected and analyzed. **(F)** Box plot of the two groups.

section of Pairwise sample analysis (**Supplementary Video S4** and **Figures 3A,B**). In the Overlapping Clonotype Frequency scatter plots, the distribution of the shared clonotypes from the selected pair of timepoint datasets deviated significantly from the diagonal. The coefficient of determination was only 0.001 between the two timepoints (**Figure 3A**). Furthermore,

a lot of high frequency clonotypes were found in the fifth timepoint TCR sequencing data of WDJ patient from the Un-Overlapping Clonotype Frequency scatter plots (**Figure 3B**). The differences between the pair of TCR sequencing data is useful as a comparison between extremes in this demonstration (since there are additional timepoints), but may just as readily serve as the primary analysis of interest in alternative study designs.

Multi-Sample Analysis

The section of Multi-samples Analysis provides a number of statistical analysis methods that are categorized into Description Statistics of TCR clonotypes, Similarity Statistical analysis between grouped datasets, and Biodiversity Statistical analysis of grouped datasets. The Description Statistics of TCR clonotypes was executed with pre-defined experimental factors Time_point in the WDJ Experiment Design Files (**Supplementary Video S5** and **Figures 3C,D**). In Clonal space homeostasis analysis, it was shown that the proportional distribution of the fourth timepoint TCR clonotypes differed from other timepoint (**Figure 3C**). In Clonotype Tracking analysis, the change of the high frequency TCR clonotypes from five timepoint demonstrated that the $CD8^+$ T cells of WDJ patient were expanded in second timepoint and contracted in third and fourth timepoint, then expanded in fifth timepoint again (**Figure 3D**). However, these types of visualizations can also be easily applied to explore the flow of T cell clones between different tissues, and each group can also be readily reordered to help facilitate ease of comprehension.

The statistical analysis on the similarity index and diversity index of TCR sequencing dataset also is developed in the VisTCR. For instance, the Bio-diversity index analysis calculated the diversity index of the TCR sequencing data according to factors set in the Experiment Design File (**Supplementary Video S6** and **Figure 3E**). In Pairwise Diversity Analysis, it was found that the diversity index (Shannon entropy) of DRESS patients was significantly lower than healthy donors ($p < 0.005$, Wilcoxon Test). The lower diversity of DRESS patients is consistent with the expected expansion of antigen specific CD8 + T cells (**Supplementary Video S6** and **Figure 3F**).

Applicability of visTCR on Mouse Data

To further demonstrate the easy and general applicability of VisTCR, we also provide an additional worked example using a publicly available mouse tumor TCRseq dataset with a distinct experimental design (Aoki et al., 2018). Simple visualization of clonal homeostasis and Shannon diversity in the peripheral blood, tumor, and draining lymph node samples yielded the expected result of the tumor samples having lowered diversity and more highly expanded clones (**Supplementary Figure 2A**). Pairwise analysis of the blood and lymph node samples was similarly consistent with the reported results, and offered a simple statistical test for significance (**Supplementary Figures 2B–E**). Additional clustering and correlation across the three sample types considered could also be easily performed in VisTCR. The frequency of the dominant clone in the tumor samples could also be readily recovered and traced across the other samples. Taken together, VisTCR make it easier for users to perform their standard and unique analysis tasks.

Additional Human Data Analysis of Sezary Syndrome

As an additional test case of the consistency of the VisTCR data analyses, we further replicated our workflow on a published dataset of peripheral blood samples from patients with Sezary syndrome, a form of cutaneous T cell lymphoma (Ruggiero et al., 2015). Consistent with the published results, the patients with Sezary syndrome showed more limited usage of TRBV chains compared to healthy controls (**Supplementary Figures 3A,B**). We could also observe that the Sezary patients had hyperexpansion of a number of clonotypes, with spectratyping showing a sharp dropoff in the detection of smaller clones as compared to healthy controls (**Supplementary Figure 3C-D**). These samples had lower performance in diversity metrics as a consequence (**Supplementary Figures 3D,E**). Taken together, these results generated using our analysis tool are qualitatively consistent with those generated using other utilities. VisTCR may thus also be useful for quickly performing third-party data re-analysis.

CONCLUSION

VisTCR has been developed to parse, evaluate, and statistically analyze the TCR repertoire data with a user-friendly GUI. The data management module provides simple functions to organize the TCR sequencing data, and the data analysis module integrates most of the popular methods for TCR repertoire analysis with an intuitive analysis workflow. We believe that VisTCR may help make TCR repertoire analysis more accessible to wet-lab scientists, and help unlock the full potential of TCRseq data.

DATA AVAILABILITY STATEMENT

The open source code of VisTCR is available for free public download at the GitHub repository: https://github.com/qingshanni/VisTCR.Publicly available datasets were analyzed in this study. These data can be found here: SRA (PRJNA611474 and PRJNA287162) and GEO (GSE115425).

ETHICS STATEMENT

Ethical review and approval was not required for this study because this study only involved re-analysis of published and publicly available datasets that had been previously approved and does not require further review as per institutional requirements. Original approval for the datasets used can be found in the papers referenced for each datasets cited.

AUTHOR CONTRIBUTIONS

Q-JL and YW designed the study. QN, JZ, and ZZ wrote the software code and prepared the figures. GC, LC, JG, HY, DZ, and YZ tested the function of the software. QN, JZ, ZZ, Q-JL, and YW

wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National key project of china (Grant No. 2016YFA0502200), China Postdoctoral Science Foundation Funded Project (Grant No. 2015M582843), and Basic Science and Frontier Technology Research Project of Chongqing (cstc2017jcyjAX0198). JG was supported by a fellowship grant from the Sigrid Juselius Foundation.

ACKNOWLEDGMENTS

We thank Qingzhu Jia, Xuezhong Yu, and Ning Jiang for their thoughtful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00771/full#supplementary-material

FIGURE S1 | The GUI of clone extract methods used in VisTCR. Three major methods are included and can be chosen by the user as follows; **(A)** Decombinator **(B)** MiTCR **(C)** MiXCR.

FIGURE S2 | Example analysis of mouse tumor dataset by using VisTCR. **(A)** Clonal homeostasis analysis. **(B)** Bio-diversity analysis by using Shannon diversity index. **(C–E)** Pairwise diversity analysis.

TABLE S1 | TCR sequencing data analysis methods in VisTCR software.

FILE S1 | Experiment design file for analyzing all 5 $CD8^+$ samples from DRESS patient WDJ. The experiment design file is used to define the specific experimental conditions and any dependent variables or factors that can be used in TCR repertoire data analysis.

FILE S2 | Experiment design file for analyzing samples from 8 DRESS patients and 8 healthy donors.

VIDEO S1 | Uploading the sequencing data files into Data Storage Module. This video displays the experimental data management functions provided by Data Storage Module in VisTCR. Firstly, an experiment is created with title and description. Then, the raw TCR sequencing data belonging to the experiment are uploaded one by one. Finally, the quality of raw sequencing data is checked.

VIDEO S2 | Creating an analysis task in the Data Analysis Module. Firstly, experiment design files are created by using Notepad ++, and saved in the CSV format. Then, a new analysis project is created by using wizard mode in VisTCR. In this process, the project title and description is set, the method for parsing raw TCR sequencing data is selected, and the experiment design file created previously is uploaded.

VIDEO S3 | Single sample analysis in VisTCR. This video displays single sample analysis functions provided by Data Analysis Module in VisTCR, including their TRBV and/or TRBJ usage, CDR3 spectratype, and their clonotype distribution.

VIDEO S4 | Pairwise sample analysis in VisTCR. This video displays pairwise sample analysis functions provided by Data Analysis Module in VisTCR, including samples selection, overlapping and un-overlapping clonotype distribution and convergence analyses.

VIDEO S5 | Description statistics analysis in VisTCR. This video displays description statistics analysis functions provided by Data Analysis Module in VisTCR, including most abundant clonotypes, clonal space homeostasis, clonotype tracking, overlap analysis.

VIDEO S6 | Multi-sample analysis of DRESS patients and healthy donors. This video displays some multi-sample analysis functions used to analyze DRESS patients and healthy donors, including most abundant clonotypes, clonal space homeostasis, bio-diversity index, and pairwise diversity analysis.

REFERENCES

Aoki, H., Ueha, S., Shichino, S., Ogiwara, H., Hashimoto, S., Kakimi, K., et al. (2018). TCR repertoire analysis reveals mobilization of novel CD8+ T cell clones into the cancer-immunity cycle following Anti-CD4 antibody administration. *Front. Immunol.* 2018:3185. doi: 10.3389/fimmu.2018.03185

Aris, M., Bravo, A. I., Pampena, M. B., Blanco, P. A., Carri, I., Koile, D., et al. (2018). Changes in the TCRbeta repertoire and tumor immune signature from a cutaneous melanoma patient immunized with the CSF-470 vaccine: a case report. *Front. Immunol.* 9:955. doi: 10.3389/fimmu.2018.00955

Bachle, M., and Kirchberg, P. (2007). Ruby on rails. *IEEE Softw.* 24, 105–108.

Bagaev, D. V., Zvyagin, I. V., Putintseva, E. V., Izraelson, M., Britanova, O. V., Chudakov, D. M., et al. (2016). VDJviz: a versatile browser for immunogenomics data. *BMC Genomics* 17:453. doi: 10.1186/s12864-016-2799-7

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., et al. (2015). MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12, 380–381. doi: 10.1038/nmeth.3364

Bolotin, D. A., Shugay, M., Mamedov, I. Z., Putintseva, E. V., Turchaninova, M. A., Zvyagin, I. V., et al. (2013). MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* 10, 813–814. doi: 10.1038/nmeth.2555

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE Trans. Vis. Computer Graph.* 17, 2301–2309.

Bystry, V., Reigl, T., Krejci, A., Demko, M., Hanakova, B., Grioni, A., et al. (2017). ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics* 33, 435–437. doi: 10.1093/bioinformatics/btw634

Duez, M., Giraud, M., Herbert, R., Rocher, T., Salson, M., and Thonier, F. (2016). Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS One* 11:e0166126. doi: 10.1371/journal.pone.0166126

Fahl, S. P., Coffey, F., Kain, L., Zarin, P., Dunbrack, R. L. Jr., Teyton, L., et al. (2018). Role of a selecting ligand in shaping the murine gammadelta-TCR repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1889–1894. doi: 10.1073/pnas.1718328115

Gardner, P. P., Shugay, M., Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., et al. (2015). VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.* 11:e1004503. doi: 10.1371/journal.pcbi.1004503

Han, Y., Li, H., Guan, Y., and Huang, J. (2015). Immune repertoire: a potential biomarker and therapeutic for hepatocellular carcinoma. *Cancer Lett.* 379, 206–212. doi: 10.1016/j.canlet.2015.06.022

Hou, X. L., Wang, L., Ding, Y. L., Xie, Q., and Diao, H. Y. (2016). Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes Immun.* 17, 153–164. doi: 10.1038/gene.2016.9

Jiang, Q., Zhao, T., Zheng, W., Zhou, J., Wang, H., Dong, H., et al. (2018). Patient-shared TCRbeta-CDR3 clonotypes correlate with favorable prognosis in chronic hepatitis B. *Eur. J. Immunol.* 48, 1539–1549. doi: 10.1002/eji.201747327

Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A. U., Babel, N., Reinert, K., et al. (2015). IMSEQ - a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* 31, 2963–2971. doi: 10.1093/bioinformatics/btv309

Nazarov, V. I., Pogorelyy, M. V., Komech, E. A., Zvyagin, I. V., Bolotin, D. A., Shugay, M., et al. (2015). tcR: an R package for T cell receptor repertoire

advanced data analysis. *BMC Bioinformatics* 16:175. doi: 10.1186/s12859-015-0613-1

Newell, E. W., and Davis, M. M. (2014). Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat. Biotechnol.* 32, 149–157. doi: 10.1038/nbt.2783

Niu, J., Jia, Q., Ni, Q., Yang, Y., Chen, G., Yang, X., et al. (2015). Association of CD8(+) T lymphocyte repertoire spreading with the severity of DRESS syndrome. *Sci. Rep.* 5:9913. doi: 10.1038/srep09913

Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.* 25, 646–652. doi: 10.1016/j.coi.2013.09.017

Ruggiero, E., Nicolay, J. P., Fronza, R., Arens, A., Paruzynski, A., Nowrouzi, A., et al. (2015). High-resolution analysis of the human T-cell receptor repertoire. *Nat. Commun.* 6:8081. doi: 10.1038/ncomms9081

Schaller, S., Weinberger, J., Jimenez-Heredia, R., Danzer, M., Oberbauer, R., Gabriel, C., et al. (2015). ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. *BMC Bioinformatics* 16:252. doi: 10.1186/s12859-015-0687-9

Six, A., Mariotti-Ferrandiz, M. E., Chaara, W., Magadan, S., Pham, H. P., Lefranc, M. P., et al. (2013). The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front. Immunol.* 4:413. doi: 10.3389/fimmu.2013.00413

Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., and Chain, B. (2013). Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* 29, 542–550. doi: 10.1093/bioinformatics/btt004

Venturi, V., Price, D. A., Douek, D. C., and Davenport, M. P. (2008). The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* 8, 231–238. doi: 10.1038/nri2260

Yang, X., Liu, D., Lv, N., Zhao, F., Liu, F., Zou, J., et al. (2014). TCRklass: a new K-string-based algorithm for human and mouse TCR repertoire characterization. *J. Immunol.* 194, 446–454. doi: 10.4049/jimmunol.1400711

Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41, W34–W40. doi: 10.1093/nar/gkt382

Conflict of Interest: ZZ was employed by Biowavelet Ltd., Chongqing, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership